Human Mutation

Meiotic recombination favors the spreading of deleterious mutations in human populations

Journal:	*Human Mutation*
Manuscript ID:	humu-2010-0316.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	01-Oct-2010
Complete List of Authors:	Necsulea, Anamaria; Université Claude Bernard Lyon 1, Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558; Université de Lausanne, Centre Intégratif de Génomique Duret, Laurent; Université Claude Bernard Lyon 1 , Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558 Popa, Alexandra; Université Claude Bernard Lyon 1, Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558 Cooper, David; Cardiff University, Institute of Medical Genetics, College of Medicine Stentson, Peter; Cardiff University, Institute of Medical Genetics, College of Medicine Cardiff Mouchiroud, Dominique; Université Claude Bernard Lyon 1, Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558 Gautier, Christian; Université Claude Bernard Lyon 1, Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558
Key Words:	human disease-associated mutations, meiotic recombination, GC-biased gene conversion, human polymorphisms, derived allele frequencies, non-synonymous mutations

Title page

Meiotic recombination favors the spreading of deleterious mutations in human populations

Anamaria Necşulea[1], Alexandra Popa[1], David N. Cooper[2], Peter D. Stenson[2], Dominique Mouchiroud[1], Christian Gautier[1], Laurent Duret[1]

[1] Université de Lyon; Université Lyon 1; CNRS; UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 43 Boulevard du 11 Novembre 1918, Villeurbanne F-69622, France

[2] Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK

Corresponding author:

Laurent Duret
Laboratoire de Biométrie et Biologie Evolutive
Université Claude Bernard - Lyon 1
43 Boulevard du 11 Novembre 1918,
Villeurbanne F-69622, France
Telephone: +33 4 72 44 62 97
FAX: +33 4 72 43 13 88
E-mail: laurent.duret@univ-lyon1.fr

Running head: Recombination and disease-associated mutations

Abstract

Although mutations that are detrimental to the fitness of organisms are expected to be rapidly purged from populations by natural selection, some disease-causing mutations are present at high frequencies in human populations. Several non-exclusive hypotheses have been proposed to account for this apparent paradox (high new mutation rate, genetic drift, overdominance or recent changes in selective pressure). However, the factors ultimately responsible for the presence at high frequency of disease-causing mutations are still contentious. Here we establish the existence of an additional process that contributes to the spreading of deleterious mutations: GC-biased gene conversion (gBGC), a process associated with recombination which tends to favor the transmission of GC-alleles over AT-alleles. We show that the spectrum of amino-acid altering polymorphisms in human populations exhibits the footprints of gBGC. This pattern cannot be explained in terms of selection and is evident with all non-synonymous mutations, including those predicted to be detrimental to protein structure and function, and those implicated in human genetic disease. We present simulations to illustrate the conditions under which gBGC can extend the persistence time of deleterious mutations in a finite population. These results indicate that gBGC meiotic drive contributes to the spreading of deleterious mutations in human populations.

Keywords: human disease-associated mutations; meiotic recombination; GC-biased gene conversion; non-synonymous mutations; polymorphisms; derived allele frequencies

Deleted: M
Deleted: . Somewhat paradoxically
Deleted: that cause genetic
Deleted:
Deleted: disease
Deleted: unclear and hotly
Deleted: debated
Deleted: . gBGC
Deleted: is
Deleted: Here
Deleted: w
Deleted: known to be
Deleted: ,
Deleted: ,
Deleted: ,
Deleted: human
Deleted: ,

Introduction

The majority of disease-causing mutations (DMs) detected in human populations are very recent, having only been transmitted over a few generations at most [Slatkin and Rannala, 2000]. A substantial fraction of DMs nevertheless correspond to more ancient mutations that have persisted for a large number of generations. Several non-exclusive hypotheses have been proposed to explain why such detrimental mutations could have escaped negative selection. First, detrimental mutations that have a limited impact on reproductive success (e.g. mutations causing late-onset diseases) can spread simply by genetic drift [Kryukov et al., 2007] Second, some DMs confer a selective advantage upon heterozygotes (overdominance) [Dean et al., 2002]. Third, some DMs may have attained a high population frequency in the past because they were once advantageous under environmental conditions that no longer pertain [Di Rienzo and Hudson, 2005]. Finally, some DMs may occur at high frequency because of a high *de novo* mutation rate or a germ-line selective advantage [Choi et al., 2008].

Population genetic models indicate that in addition to genetic drift and natural selection, there is a third process that can contribute to the spreading of mutations within a population: biased gene conversion (BGC). Gene conversion occurs during homologous recombination and involves the non-reciprocal transfer of sequence information between the two recombining DNA molecules. This process is said to be biased if one of the two DNA molecules involved is more likely than the other to be the donor. Gene conversion can affect paralogous sequences duplicated in the genome or different alleles at a given locus [Chen et al., 2007]. In the case of allelic gene conversion, BGC leads to an excess of the 'favored' allele in the pool of gametes and hence tends to increase the frequency of this allele in the population. Theoretical analyses have shown that, as with selection, BGC can increase the probability of fixation of the favored allele [Nagylaki, 1983].

Deleted: (
Deleted:)
Deleted: ing
Deleted: (
Deleted: ,
Deleted:)
Deleted: .
Deleted: (
Deleted: ,
Deleted:)
Deleted: cquired
Deleted: (
Deleted:)
Formatted: Font: Italic
Deleted: (
Deleted: ,
Deleted:)
Deleted: besides
Deleted: (Chen, et al., 2007)
Deleted: (
Deleted:)

Although the theoretical consequences of the BGC process have been known for some time, the potential practical importance of this phenomenon has remained largely unstudied. Recently, the analysis of polymorphism and nucleotide substitution patterns in primates has provided firm evidence for BGC acting genome-wide, favoring GC alleles over AT alleles (for a review, see Duret and Galtier, 2009a). Indeed, this process of GC-biased gene conversion (gBGC) appears to be the major determinant of the evolution of base composition at silent sites (non-coding regions, synonymous codon positions) in primate genomes [Duret and Arndt, 2008]. Further, there is now good evidence that gBGC has impacted upon the evolution of functional sequences, both in regulatory non-coding sequences [Duret and Galtier, 2009b; Galtier and Duret, 2007] and in protein-coding exons [Berglund, et al., 2009; Galtier, et al., 2009]. Importantly, these results indicate that, in our species' evolutionary past, gBGC is likely to have hampered the action of purifying selection and led to the fixation of deleterious mutations.

Here we have sought to determine whether gBGC influences the frequency of deleterious non-synonymous polymorphisms in extant human populations. To this end, we investigated the segregation patterns of AT→GC and GC→AT single nucleotide polymorphisms (SNPs) according to the local recombination rate. We also analyzed different classes of non-synonymous SNPs, predicted to be deleterious or known to be involved in genetic disease, using synonymous and non-coding SNPs as a neutral control. All classes of SNPs were found to display the hallmarks of the gBGC process. Further, we provide evidence that these segregation patterns cannot be explained by ascertainment bias in SNP detection, artifacts in SNP orientation, or other biological processes such as natural selection. In support of these observations, we present simulations to illustrate the conditions under which gBGC can extend the persistence of deleterious mutations in finite populations. We conclude that gBGC

Deleted: (
Deleted:)
Deleted: (
Deleted:)
Deleted: (
Deleted:)
Deleted: (
Deleted:)
Deleted: aim
Deleted: For that purpose
Deleted: s
Deleted: as well as
Deleted: We find that
Deleted: a
Deleted: show
Deleted: W
Deleted: by
Deleted: in

has not only had a substantial impact on human evolution but is also highly relevant to human health and disease.

Material and methods

Single nucleotide polymorphism data. To determine the frequency of SNPs in human populations, we used the data gathered in the HapMap Project phase III, release 27 [Frazer et al., 2007]. We analyzed data from four HapMap populations: YRI (Yoruba in Ibadan, Nigeria), JPT (Japanese in Tokyo), CHB (Han Chinese in Beijing) and CEU (Utah residents with ancestry from northern and western Europe) and we grouped the CHB and JPT samples into a single set. We analyzed only SNPs that were polymorphic in the unrelated individuals genotyped in each sample (3,566,377 total, Supp. Table S1). Ensembl annotations [Hubbard et al., 2009] were used to determine the positions of SNPs with respect to transcripts and coding sequences. Four classes of polymorphism were retained for analysis: intergenic, intronic, protein-coding synonymous and protein-coding non-synonymous.

As a complement, we used an independent polymorphism dataset comprising 39,440 autosomal SNPs, found exclusively in coding sequences, at both synonymous and non-synonymous positions [Lohmueller, et al., 2008]. These SNPs were determined by direct exon sequencing in 10,150 transcripts, for two population samples (hereafter termed AFR and CAU): 15 African-American individuals (30,718 SNPs) and 20 European-American individuals (22,514 SNPs, Supp. Table S2).

Inference of ancestral and derived alleles. We determined the ancestral and derived states of human polymorphisms using human–chimpanzee whole-genome alignments, obtained from the UCSC Genome Browser [Rhead et al., 2010] through Galaxy [Giardine et al., 2005].

Deleted: ,
Deleted: and
Deleted: a
Deleted: issue for
Deleted: ----------Page Break----------
Deleted: (
Deleted: ,
Deleted:)
Deleted: (
Deleted: ,
Deleted:)
Deleted: s
Deleted: (
Deleted:)
Deleted:
Deleted:
Field Code Changed
Deleted: (
Formatted: Font: Times New Roman
Deleted: ,
Formatted: Font: Times New Roman
Deleted:)
Deleted: (
Deleted: ,
Deleted:)

To infer the most likely ancestral and derived alleles for each SNP, we used a maximum likelihood approach that takes into account the hypermutability of CpG dinucleotides [Duret and Arndt, 2008]. Starting from whole-genome alignments of the human and chimpanzee sequences, we constructed triple alignments that included two sequences for the human population, corresponding to the two alleles observed for each SNP. The allocation of alleles to the two human sequences was performed randomly. We then inferred the ancestral sequence for the human population, thereby obtaining for each genomic position a probability distribution for the identity of the ancestral nucleotide. The ancestral nucleotide was randomly drawn according to these four probabilities. In our analysis, we included only SNPs with a constant 5'-3' context (*i.e.* positions with two neighboring SNPs were removed, and we required that the human and chimpanzee nucleotides should be identical).

To confirm that this first approach had not been misled by ancestral 'misinference' issues, we also used a second approach, developed by Hernandez, et al. [2007a], which corrects the spectrum of derived allele frequencies, obtained by parsimonious reasoning, using a context-dependent model of sequence evolution (software kindly provided by Ryan D. Hernandez). We only considered SNPs found within a constant 5'-3' context, as defined above. As indicated by the authors, we further restricted our dataset to positions where the chimpanzee nucleotide corresponded to one of the two alleles observed in the human population. The context-dependent site frequency spectrum obtained by maximum parsimony was then corrected using the model proposed by Hernandez, et al. [2007a].

As noted previously [Gibbs, et al., 2007], for disease-associated mutations, the disease-associated allele sometimes represents the ancestral state; here, we focused exclusively on SNPs for which the derived allele was associated with the disease.

SNP sampling and derived allele frequency spectrum

Deleted: (
Deleted:)
Deleted: (
Deleted: ,
Deleted:)
Deleted: consists in
Deleted: ing
Deleted: by
Deleted: (
Deleted: ,
Deleted:)
Deleted: (
Deleted: ,
Deleted:)
Deleted: ,

The number of genotyped chromosomes varies widely between individual SNPs. The correction method developed by Hernandez et al. [2007a] requires the derived allele frequency spectrum to be constructed employing the same number of chromosomes for all SNPs. To fulfill this requirement, we applied the following procedure (as proposed by Hernandez et al. [2007b]): we computed the minimum number of sampled chromosomes (n_{min}) for a given SNP dataset and then estimated the derived allele frequencies for a dataset reduced to n_{min} chromosomes. For a SNP that was originally present in n out of m sampled chromosomes, the probability that it will be present at a frequency i in the reduced sample is given by the hypergeometric distribution: $\frac{C_n^i \times C_{m-n}^{n_{\min}-i}}{C_m^{n_{\min}}}$, where C_u^v is the number of choices of v elements among u. Using this formula, we can generate the expected derived allele frequency spectrum in a subsample of n_{min} chromosomes. Note that this procedure was applied independently for each class of SNPs analyzed here (intergenic, intronic, synonymous SNPs, etc.). The n_{min} values for each SNPs sample and for each region are given in Supp. Table S5.

Recombination rates and hotspots. The positions of 34,136 recombination hotspots were taken from HapMap release 21 [Myers et al., 2005], and converted from hg17 to hg18 assembly coordinates using the *liftover* utility from the UCSC Genome Browser [Rhead et al., 2010]. We also computed the regional recombination rates in 10kb sliding windows for autosomal mutations using the genetic maps provided by [Frazer et al., 2007], release 36.

Disease-associated mutations. We extracted 45,751 disease-associated mutations occurring in protein-coding sequences from HGMD release 2008.3 [Stenson et al., 2009]. Using annotations from the Ensembl database [Hubbard et al., 2009], release 49, we were able to

Deleted: (
Deleted: ,
Deleted: ,
Deleted:)
Deleted:
Deleted: (
Deleted: ,
Deleted: ,
Deleted:
Deleted:)
Deleted: ,
Deleted: (
Deleted: ,
Deleted:)
Deleted: (
Formatted: Font: Times New Roman
Deleted: ,
Field Code Changed
Formatted: Font: Times New Roman
Deleted:)
Deleted: (
Deleted: ,
Deleted:)
Deleted: (
Deleted: ,
Deleted:)
Deleted: (
Deleted: ,
Deleted:)
Deleted: ,

map unambiguously onto the human genome the positions of 43,953 disease-associated mutations. 193 mutations were synonymous and hence were excluded - here we only analyzed non-synonymous mutations (34,814 missense and 8,946 nonsense).

HGMD mutations are allocated to four distinct classes with respect to their association with disease: DM, mutations regarded as being a direct cause of disease; DP, polymorphisms exhibiting a significant statistical association with disease but without additional functional evidence supporting their involvement; DFP, disease-associated polymorphisms with additional functional evidence supporting their direct involvement; FP, polymorphisms reported to affect the structure, function or expression of the gene (or gene product), but with no known disease association (Supp. Table S3).

PolyPhen predictions. To predict which non-synonymous SNPs present in HapMap are potentially damaging for protein structure and function, we used PolyPhen predictions for dbSNP build 126 [Sunyaev et al., 2001]. For the exon sequencing dataset, we used the PolyPhen predictions provided by the authors [Lohmueller et al., 2008] (Supp. Table S4). We focused on the SNPs predicted to be "probably damaging", for which the derived allele has been shown to be the deleterious allele in 99% of cases [Lohmueller et al., 2008].

Definition of recombination classes. To define regions of high and low recombination, we sorted each SNP dataset according to the minimum distance to a recombination hotspot, and then divided the dataset into three equal-sized classes. Only the first and the third classes were compared in order to maximize the crossover rate difference between the high and low recombination regions. This procedure was applied independently for each genomic region

Deleted: ;
Deleted: For
Deleted: , there are
Deleted: of
Deleted: supporting
Deleted: the
Deleted: (
Deleted: ,
Deleted:)
Deleted: (
Deleted: ,
Deleted:)
Deleted: the
Deleted: (
Deleted: ,
Deleted:)

(intergenic, intronic, coding synonymous etc.) and for each HGMD and PolyPhen subset of SNPs.

Statistical analyses. All statistical analyses were performed with the R environment [R Development Core Team, 2008]. To test the effect of gBGC, we compared the mean derived allele frequencies (DAF) for AT→GC and GC→AT mutations. Given that the distribution of DAF is non-Gaussian, we used a randomization procedure to test the statistical significance of the mean difference (d = mean(AT→GC)-mean(GC→AT)). To do this, we randomized the direction of AT→GC and GC→AT SNPs and compared the observed d value with those obtained from 1000 randomized datasets. We computed a p-value corresponding to the proportion of simulated datasets for which the d value was higher than that observed in the real dataset; our test was thus one-tailed.

We also analyzed the difference in mean DAF between the two mutation classes (d) for regions of high and low recombination. To test if the difference in d (Δd) between the two recombination classes was statistically significant, we developed a randomization procedure: we drew randomly two sets of sites (from all possible SNPs in a given genomic region), equal in size to the original low recombination and high recombination classes, and computed Δd for the simulated dataset. A one-tailed p-value was computed by comparing the observed Δd value with 1000 simulated datasets.

Simulation of the impact of gBGC on the derived allele frequency spectrum

We used simulations to determine the expected distribution of derived allele frequencies (DAF) at loci that are subject to mutation, negative selection and biased gene conversion. The initial population was homozygous and finite following a Fisher-Wright probabilistic model

Deleted: (
Formatted: Font: Times New Roman
Field Code Changed
Formatted: Font: Times New Roman
Deleted:)
Deleted: e one
Deleted: i
Deleted: classes
Deleted: i
Deleted:

with multinomial sampling, ensuring a constant population size over time. The evolution of the derived allele frequency was simulated independently for each locus. Each simulation was performed for over 20,000 generations, at the end of which, the DAF of the derived allele was calculated.

Deleted: i
Deleted: i
Deleted: i
Deleted: i

The alleles that can segregate at each locus belong to one of two classes: S(trong) (G or C) or W(eak) (A or T). The fitness of genotypes **SS**, **SW** and **WW** are denoted respectively ω_{SS}, ω_{SW} and ω_{WW}. The mean fitness value in the population is $\bar{\omega}$:

$\bar{\omega} = z_{SS}\omega_{SS} + z_{SW}\omega_{SW} + z_{WW}\omega_{WW}$ where z denotes the zygotic frequencies.

For individuals that are heterozygous at a given locus (**SW**), we termed u the probability of conversion $S \to W$ and v the probability of conversion $W \to S$. The gene conversion bias at this site is measured through $\delta = v - u$ and has positive values when gBGC occurs. The frequency of the **S** allele is denoted p and hence the frequency of allele **W** is $1-p$. The model describes the transition from one generation, n, to the next, $n+1$, admitting panmixia, with the following equations:

adults n: f_{SS}; f_{SW}; f_{WW};

gametes n: $g_S = \frac{2f_{SS} + (1+\delta)f_{SW}}{2}$ $g_W = 1 - g_S$

zygotes $n+1$: $z_{SS} = g_S^2$; $z_{SW} = 2g_S g_W$; $z_{WW} = g_W^2$

adults $n+1$: $f^*_{SS} = \frac{\omega_{SS}}{\bar{\omega}} z_{SS}$; $f^*_{SW} = \frac{\omega_{SW}}{\bar{\omega}} z_{SW}$; $f^*_{WW} = \frac{\omega_{WW}}{\bar{\omega}} z_{WW}$

alleles $n+1$: $p_S = f^*_{SS} + \frac{1}{2} f^*_{SW}$ $p_W = 1 - p_S$

where f represents the frequency of individuals at generation n, g the frequency of gametes at generation n, and f^* the frequency of individuals at generation $n+1$.

Here we only considered mutations that are both deleterious and recessive. We termed s the selection coefficient, so that the fitness of individuals homozygous for the mutant allele is $\omega = 1 - s$. Thus, for the simulations of the fate of a newly-arisen $W \rightarrow S$ mutation in a **WW** population, we have $\omega_{SS} = \omega$ and $\omega_{SW} = \omega_{WW} = 1$ whereas for the simulations of the fate of a newly-arisen $S \rightarrow W$ mutation in an **SS** population, we have $\omega_{SS} = \omega_{SW} = 1$ and $\omega_{WW} = \omega$.

Simulations were run in populations of size N_e=10,000 with a mutation rate of 10^{-8} mutations per base-pair per individual per generation, using different combinations of gBGC coefficient (δ=0, δ=0.00013 and δ=0.0013) and selection coefficient (s=0, s=10^{-4} , s=10^{-3} and s=10^{-2}).

Deleted: te

Deleted: . And

Deleted:

Deleted:

Supporting Information

The dataset used in this publication is freely available at the following website:

ftp://pbil.univ-lyon1.fr/pub/datasets/Necsulea2010

Results

gBGC hallmarks are observed for deleterious SNPs

To investigate whether gBGC affects the segregation of deleterious mutations in human populations, we studied the spectrum of derived allele frequencies (DAFs) of non-synonymous SNPs as a function of the local recombination rate across human chromosomes. We first analyzed the HapMap dataset of human SNPs, which provides frequencies of each allele in different human populations [Frazer, et al., 2007]. We inferred the ancestral and derived alleles for SNPs by means of a maximum likelihood approach that incorporates CpG hypermutability [Duret and Arndt, 2008], using the chimpanzee genome as an outgroup. Three distinct subsets of non-synonymous polymorphisms were investigated: 1) all HapMap non-synonymous SNPs; 2) HapMap non-synonymous mutations for which the impact on the function of the protein was predicted by PolyPhen [Sunyaev, et al., 2001] to be 'probably damaging'; and 3) HapMap SNPs corresponding to disease-associated non-synonymous mutations reported in the HGMD database [Stenson et al., 2009]. We further split the HGMD dataset in order to analyze specifically those inherited mutations which are considered to be a direct cause of disease (DM), thereby excluding those mutations that have only been associated statistically with disease (Supp. Table S3). As a control, we also analyzed SNPs at silent sites, for which evidence of gBGC has already been reported [Galtier, et al., 2001; Spencer, et al., 2006; Webster and Smith, 2004]. As expected, DAFs were found to be negatively correlated with the strength of purifying selection: SNPs in non-coding regions or at synonymous codon positions exhibited the highest mean DAFs, whereas the lowest mean DAFs were observed for mutations that are known to be involved in genetic disease or that were predicted by PolyPhen to be deleterious (Fig. 1, Supp. Tables S8-S10).

The gBGC model makes two firm predictions: first, in regions of high recombination, the spectrum of derived allele frequencies (DAFs) for SNPs is expected to be skewed, with higher

Deleted: (
Deleted: ,
Deleted:)
Deleted: (
Deleted:)
Deleted: (
Deleted: ,
Deleted:)
Deleted: (
Deleted: ,
Deleted:)
Deleted: (
Deleted: ,
Deleted:)
Deleted: s

frequencies for AT→GC than for GC→AT mutations; second, this skewing is expected to be weaker in genomic regions characterized by a lower recombination rate. To test these predictions, we classified SNPs into groups of high and low recombination on the basis of their physical distance to the nearest recombination hotspot [Myers et al., 2005]; similar results were obtained when the recombination classes were computed on the basis of the average crossover rate in fixed-size sliding windows (not shown). We found that in regions of high recombination, AT→GC mutations segregated at higher frequencies than GC→AT mutations (Fig. 2, Supp. Tables S8-S10, Supp. Fig. S2-S4). This difference was statistically significant in all HapMap samples, both for silent SNPs and for the three sets of non-synonymous SNPs (Table 1). This pattern was evident even within the DM subset. For this class, the tests remained significant in only one of the HapMap samples. Nevertheless, given that our observations for the more abundant classes of mutations (silent sites, non-synonymous SNPs) were always in agreement with the gBGC hypothesis, and significantly so, the uncertainty related to the DM class is most likely only a consequence of the reduced sample size. As predicted by the gBGC model, the difference between the mean AT→GC and GC→AT frequencies is much stronger for SNPs located in regions of high recombination rate as compared to SNPs located in regions of low recombination rate (Fig. 1, Table 1, Supp. Tables S13-S15). Thus, all classes of SNPs exhibit the hallmarks of the gBGC process, not only the silent sites but also the three subsets of non-synonymous sites.

Deleted: (
Deleted: ,
Deleted:)
Deleted: a
Deleted: a
Deleted: (but
Deleted:
Deleted: , probably due to the limited amount of data available)
Deleted: .
Deleted: a
Deleted: :

Control for variations in selective pressure on non-synonymous mutations

We observed that at non-synonymous sites, GC→AT mutations segregate at lower frequency than AT→GC mutations. One potential explanation for this observation is that AT→GC non-synonymous mutations might be, on average, less deleterious than GC→AT non-synonymous mutations. To test this hypothesis, we compared AT→GC and GC→AT SNPs that lead to the

same amino-acid replacement and hence are expected to have the exact same fitness impact. In total, there are 10 amino-acid changes that can be caused both by AT→GC and GC→AT mutations. For each of the three populations, we performed pairwise comparisons of the mean DAF of AT→GC and GC→AT SNPs causing the same amino-acid changes: in 23 out of 30 comparisons, the AT→GC SNP had the highest mean DAF (Supp. Table S19). For example, the mean DAF of Q→H non-synonymous SNPs in the CEU population is 0.19 when it results from an AT→GC mutation, compared to 0.16 when it results from a GC→AT mutation. Conversely, the mean DAF of the reverse amino-acid change (H→Q) is 0.35 when it results from an AT→GC mutation, compared to 0.23 when it results from a GC→AT mutation. Thus, the mean DAF varies according to the direction of the GC-content change (AT→GC vs. GC→AT), independently of the nature of the amino-acid change. Hence, the observed differences in mean DAF between AT→GC and GC→AT non-synonymous SNPs cannot be attributed to differences in selective pressure on the corresponding amino-acid changes.

Deleted: lementary

Control for SNP ascertainment bias and ancestral misidentification

The HapMap dataset is known to be biased towards high frequency polymorphisms, and this representation bias can confound some population genetic analyses [Clark et al., 2005]. There is however no *a priori* reason why this ascertainment bias should differentially affect AT→GC- and GC→AT-derived allele frequencies. This notwithstanding, to ensure that our observations were not affected by this intrinsic bias in HapMap data, we repeated our analysis on an independent polymorphism dataset that was acquired through direct exon re-sequencing in two human populations [Lohmueller et al., 2008], and which should therefore be free of ascertainment bias. Our conclusions remained unchanged with the re-sequencing dataset: in regions of high recombination, AT→GC mutations segregated at higher frequencies than GC→AT mutations and this excess was higher than in regions of low recombination. This pattern was observed in both populations, not only for the synonymous sites but also for the

Deleted: (
Deleted: ,
Deleted:)

Deleted: (
Deleted: ,
Deleted:)

three datasets of non-synonymous sites (Table 1, Supp. Tables S11-S12, S16-S17, Supp. Fig. S5-S6). We may therefore conclude that the observed skewing of derived allele frequencies was not simply a consequence of ascertainment bias. It may be noted that the pattern appears to be stronger with the HapMap dataset as compared to the re-sequencing dataset (Table 1). By means of simulations, we showed that this difference is due to the fact that the HapMap SNP sampling strategy provides greater power to detect gBGC (see Supporting Information).

One other potential artifact that had to be considered and assessed was the possibility that the observed gBGC-like pattern stemmed from ancestral 'misinference' [Hernandez et al., 2007a]: when the mutational pattern is biased towards AT, and most notably in the case of strong context dependence (such as CpG dinucleotide mutational hotspots in mammalian genomes), maximum parsimony tends to incorrectly ascribe directionality for GC→AT mutations, yielding an apparent excess of high-frequency AT→GC SNPs [Hernandez et al., 2007a]. Nevertheless, we are confident that this artifact has not influenced our results for the following reasons. First, instead of using parsimony-based reasoning, we determined SNP directionality using a maximum-likelihood approach that takes CpG hypermutability into account [Duret and Arndt, 2008]. Second, our conclusions were unchanged when CpG sites were excluded (Supp. Table S7). Third, we repeated our analyses using the context-dependent model proposed by Hernandez and colleagues [Hernandez et al., 2007a] to correct for potential ancestral allele misidentification. With this method, the results remained in agreement with our previous observations (Supp. Table S6). Finally, it should be highlighted that the difference between the mean DAFs of AT→GC and GC→AT mutation was found to be much stronger in regions of high recombination (Fig. 1). This observation, which is consistent with the gBGC model, cannot be explained by an ancestral misinference artifact. Indeed, the pattern of substitution is more biased toward AT in regions of low recombination as compared to regions of high recombination [Duret and Arndt, 2008]. Thus, an artifactual

Deleted: can
Deleted: ic
Deleted: a
Deleted: (
Deleted: ,
Deleted:)
Deleted: s
Deleted: (
Deleted: ,
Deleted:)
Deleted: affected
Deleted: ,
Deleted: (
Deleted:)
Deleted: (
Deleted: ,
Deleted:)
Deleted: the
Deleted: (
Deleted:)
Deleted: the

increase in AT→GC DAFs caused by ancestral misinference would be expected to be stronger in regions of low recombination, in contradistinction to our own observations (Fig. 1).

Simulation of the impact of gBGC in a finite population

To investigate the impact of gBGC on the fate of deleterious mutations (AT→GC or GC→AT), we performed simulations in a finite population (effective population size N_e=10,000), considering recessive mutations subject to different selection coefficients (s) and gBGC coefficients (δ; see methods). The population-scale gBGC coefficient ($N_e\delta$) in the human genome was estimated by Spencer et al. [2006] by analyzing the DAF spectra of non-coding SNPs. In genomic regions of high recombination (defined as the top 20% of the genome with the highest recombination rate; average crossover rate= 2.5 cM/Mb) their estimate was $N_e\delta$=0.325. Given that, in the human genome, recombination is essentially confined to hotspots (typically less than 2 kb long) with an average crossover rate of about 40 cM/Mb [Myers et al., 2006], it is expected that the gBGC coefficient should be about 16 times higher in these hotspots. Recombination hotspots vary in intensity [Myers et al., 2006]. We therefore considered two values for the population-scale gBGC coefficient: $N_e\delta$=1.3 (for a moderate recombination hotspot) and $N_e\delta$=13 (for a more intense recombination hotspot).

With gBGC parameters corresponding to those of a moderate human recombination hotspot, the impact of gBGC on the DAF spectrum was clearly detectable for both nearly-neutral ($|N_es|$ = 1) and mildly deleterious mutations ($|N_es|$ = 10): compared to a situation without gBGC ($N_e\delta$=0), AT→GC segregate at higher frequency, whereas GC→AT segregate at lower frequency (Fig. 3). For the more intense recombination hotspots, the impact of gBGC on the DAF spectrum was detectable even for highly deleterious mutations ($|N_es|$ = 100).

Deleted: is
Deleted: which is
Deleted: with
Deleted: ,
Deleted: (
Deleted:)
Deleted: occurs
Deleted: in
Deleted: ,
Deleted: (
Deleted: ,
Deleted:)
Deleted: (
Deleted: ,
Deleted:)
Deleted: ,

Recombination hotspots occupy only a small fraction of the genome: among the non-synonymous SNPs that we analyzed, 6% were located within 2kb of the center of a recombination hotspot. Thus, only a limited fraction of SNPs is expected to be affected by gBGC. This explains why the skewing observed in real data (Fig. 2) is intermediate between the patterns obtained in simulations corresponding to moderate hotspots ($N_e\delta$=1.3) or to the absence of gBGC ($N_e\delta$=0) (Fig. 3). Thus, the pattern observed with real data appears to be compatible with the hypothesis that the skew in the DAF spectrum is due to gBGC affecting deleterious mutations in recombination hotspots. It should be noted that the location of recombination hotspots is extremely dynamic [Baudat et al., 2010; Myers et al. 2010], which suggests that the fraction of SNPs that are at some time affected by gBGC, might be larger than that estimated above. To obtain a more realistic estimation of the expected DAF spectra, it would be necessary to take into account not only the intensity recombination hotspots but also their dynamics.

Deleted: a

Deleted: (

Formatted: Font: Times New Roman

Deleted: ,

Field Code Changed

Deleted:

Deleted: ,

Deleted:)

Formatted: Font: Times New Roman

Deleted: hus, t

Discussion

We have shown that all functional classes of SNPs, including non-synoynmous SNPs known to be implicated in human disease, and non-synonymous SNPs predicted to be damaging for protein structure and function, exhibit the hallmarks of gBGC: the derived allele frequency of AT→GC mutations is higher than that of GC→AT mutations, and this is more promounced in regions characterized by high recombination rates. Importantly, we demonstrated that the observed excess of high-frequency SNPs in regions of high recombination does not result from sampling biases nor from artifacts of SNP directionality determination.

Deleted: ed

Deleted: so

Deleted: with

Deleted: showed

Is gBGC the only possible explanation for these observations? One alternative hypothesis to explain the fact that non-synonymous GC→AT mutations segregate at lower frequency than AT→GC mutations is that GC→AT mutations could be more deleterious that the AT→GC mutations. For instance, it has been recently shown that GC→AT mutations at hypermutable CpG sites within coding regions are under stronger purifying selection than other non-synonymous mutations [Schmidt et al., 2008]. Several observations however argue against this hypothesis. First, we note that our conclusions remained unchanged when SNPs occurring within a CpG context were excluded (Supp. Table S7). Second, comparison of GC→AT and AT→GC mutations causing the same amino-acid changes confirmed that the higher mean DAF of the latter cannot be attributed to a weaker impact on the encoded protein. Moreover, this hypothesis that AT→GC mutations are relatively less deleterious cannot explain why their mean DAF increases with the recombination rate. Finally, we have shown that the DAF pattern is consistent over all classes of SNP, including those located in intergenic and intronic regions, which may be presumed to be largely free of selective pressure. It has been previously demonstrated that the relationship between recombination and the evolution of GC-content in non-coding regions is the consequence of gBGC and not selection [Duret and

Deleted: But

Deleted: i

Deleted: (

Deleted: ,

Deleted:)

Deleted: however

Deleted: W

Deleted: however

Deleted: the

Deleted: not

Deleted: (

Arndt, 2008]. Hence, the most parsimonious explanation for our findings is that both silent sites and non-synonymous sites are subject to gBGC.

Taken together, the data presented are consistent with the hypothesis that biased gene conversion is responsible for the excess of AT→GC SNPs segregating at high frequency in regions of high recombination. This result has important implications for human health because it indicates that recombination, via gBGC, leads to an increase in the frequency of disease-causing AT→GC mutations in human populations. It should be stressed that the impact of gBGC on deleterious mutations is not always negative. Indeed, a majority (58.7%) of known DMs correspond to GC→AT mutations. Thus, for a majority of DMs, gBGC acts in such a way as to limit their probability of spreading. However, the price to pay for this positive influence of gBGC is that it can lead to an increase in the frequency of disease-causing AT→GC mutations in human populations. We speculate that the genes most likely to be influenced by this effect will be those that are AT-rich (*i.e.* for which there are more opportunities for AT→GC mutations) and which coincide with recombination hotspots: an additional argument for these hotspots being an Achilles' heel of the human genome [Duret and Galtier, 2009b; Galtier and Duret, 2007].

Deleted:)

Deleted: implications

Deleted: literally

Deleted: (

Deleted:)

Acknowledgements

This work was supported by the Centre National de la Recherche Scientifique and by the Agence Nationale de la Recherche (ANR-08-GENO-003-01). All computations were performed at the IN2P3 Computing Center. We thank Nicolas Galtier, Sylvain Glémin, Alex Kondrashov and one anonymous referee for their helpful comments, Rasmus Nielsen and Dara Torgersson for providing us with SNP data and Ryan D. Hernandez for providing the software used to correct the DAF spectrum.

Deleted: ,

Deleted: done

Deleted: c

Deleted: c

Deleted: and

Deleted:

Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, de Massy B. 2010. *PRDM9* is a major determinant of meiotic recombination hotspots in humans and mice. Science 327:836-840.

Berglund J, Pollard KS, Webster MT. 2009. Hotspots of biased nucleotide substitutions in human genes. PLoS Biol 7:e26.

Chen JM, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP. 2007. Gene conversion: mechanisms, evolution and human disease. Nat Rev Genet 8:762-75.

Choi SK, Yoon SR, Calabrese P, Arnheim N. 2008. A germ-line-selective advantage rather than an increased mutation rate can explain some unexpectedly common human disease mutations. Proc Natl Acad Sci USA 105:10143-10148.

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15:1496-1502.

Dean M, Carrington M, O'Brien SJ. 2002. Balanced polymorphism selected by genetic versus infectious human disease. Annu Rev Genomics Hum Genet 3:263-292.

Di Rienzo A, Hudson RR. 2005. An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends Genet 21:596-601.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet 4:e1000071.

Duret L, Galtier N. 2009a. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet 10:285-311.

Duret L, Galtier N. 2009b. Comment on "Human-specific gain of function in a developmental enhancer". Science 323:714; author reply 714.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM and others. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-861.

Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet 23:273-277.

Galtier N, Duret L, Glemin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends Genet 25:1-5.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159:907-911.

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J and others. 2005. Galaxy: a platform for interactive large-scale genome analysis. Genome Res 15:1451-1455.

Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK and others. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222-234.

Hernandez RD, Williamson SH, Bustamante CD. 2007a. Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol Biol Evol 24:1792-1800.

Hernandez RD, Williamson SH, Zhu L, Bustamante CD. 2007b. Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-content in humans. Mol Biol Evol 24:2196-2202.

Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L and others. 2009. Ensembl 2009. Nucleic Acids Res 37(Database issue):D690-697.

Kryukov GV, Pennacchio LA, Sunyaev SR. 2007. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet 80:727-739.

Formatted: Font: Times New Roman
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman, Italic
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman
Deleted: (5967)
Formatted: Font: Times New Roman
Formatted: Font: Times New Roman
Formatted ... [1]
Deleted: (1)
Deleted: (10)
Deleted:
Deleted:
Deleted: (29)
Deleted: (11)
Deleted: (11)
Deleted: (5)
Deleted: (5915)
Deleted: (7164)
Deleted: (6)
Deleted: (1)
Deleted: (2)
Deleted: (10)
Deleted: (5822)
Deleted: (8)
Deleted: (10)
Deleted: (4)

Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R and others. 2008. Proportionally more deleterious genetic variation in European than in African populations. Nature 451:994-997.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310:321-324.

Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G, Donnelly P. 2010. Drive against hotspot motifs in primates implicates the *PRDM9* gene in meiotic recombination. Science 327:876-879.

Myers S, Spencer CC, Auton A, Bottolo L, Freeman C, Donnelly P, McVean G. 2006. The distribution and causes of meiotic recombination in the human genome. Biochem Soc Trans 34:526-530.

Nagylaki T. 1983. Evolution of a finite population under gene conversion. Proc Natl Acad Sci USA 80:6278-6281.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ and others. 2010. The UCSC Genome Browser database: update 2010. Nucleic Acids Res 38(Database issue):D613-619.

Schmidt S, Gerasimova A, Kondrashov FA, Adzhubei IA, Kondrashov AS, Sunyaev S. 2008. Hypermutable non-synonymous sites are under stronger negative selection. PLoS Genet 4:e1000281.

Slatkin M, Rannala B. 2000. Estimating allele age. Annu Rev Genomics Hum Genet 1:225-249.

Spencer CC, Deloukas P, Hunt S, Mullikin J, Myers S, Silverman B, Donnelly P, Bentley D, McVean G. 2006. The influence of recombination on human genetic diversity. PLoS Genet 2:e148.

Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN. 2009. The Human Gene Mutation Database: 2008 update. Genome Med 1:13.

Sunyaev S, Ramensky V, Koch I, Lathe W, 3rd, Kondrashov AS, Bork P. 2001. Prediction of deleterious human alleles. Hum Mol Genet 10:591-597.

Webster MT, Smith NG. 2004. Fixation biases affecting human SNPs. Trends Genet 20:122-126.

Deleted: (7181)

Deleted: (5746)

Formatted ... [2]

Deleted: (5967)

Formatted ... [3]

Deleted: (Pt 4)

Deleted: … …(20) ... [4]

Formatted ... [5]

Formatted ... [6]

Deleted: (11)

Deleted: (9)

Deleted: (1)

Deleted: (6)

Deleted: (3)

Figure legends

Figure 1. Mean derived allele frequencies for AT→GC and GC→AT alleles in regions of high and low recombination, for the HapMap YRI sample, for different genomic regions and classes of non-synonymous SNPs. Dark gray: AT→GC, light gray: GC→AT mutations. Solid bars: low recombination, hatched bars: high recombination. Probably damaging: HapMap non-synonymous SNPs predicted by Polyphen to be probably damaging. HGMD: entire HGMD dataset. DM: inherited mutations known to be a direct cause of disease (HGMD mutations minus those that have only been associated statistically with disease).

Figure 2. Derived allele frequency spectra for the HapMap YRI sample, for different genomic regions and classes of non-synonymous SNPs. The data presented here relate only to the high recombination class. Dark gray: AT→GC mutations, light gray: GC→AT mutations.

Figure 3. Derived allele frequency spectrum obtained through simulations with different parameter sets. Represented in light gray are the distributions of derived allele frequencies for GC→AT alleles, and in dark gray, those of AT→GC alleles. The population-scaled selection coefficient (Nes) and the population-scaled biased gene conversion parameter (Ne δ) is indicated for each graph.

Deleted: ies

Formatted: Normal, Left

Deleted: Table legends¶
Table 1: Summary table for the BGC hallmarks for the HapMap and resequencing SNP datasets. The difference in mean derived allele frequencies between AT→GC and GC→AT SNPs is denoted by d. d_H is the value of d in regions of high recombination. Δd represents the difference in d between the high and low recombination regions. Dark green: values are positive and significantly different from zero, with a p-value < 0.05. Light green: values are positive but not significantly different from zero. Light red: values are negative but not significantly different from zero. No cases were found where d_H or Δd were significantly lower than zero.

Page 21: [1] Formatted INSRV 10/1/2010 8:16:00 AM
Font: Times New Roman

Page 22: [2] Formatted INSRV 10/1/2010 8:21:00 AM
Font: Times New Roman

Page 22: [2] Formatted INSRV 10/1/2010 8:21:00 AM
Font: Times New Roman

Page 22: [2] Formatted INSRV 10/1/2010 8:21:00 AM
Font: Times New Roman

Page 22: [2] Formatted INSRV 10/1/2010 8:21:00 AM
Font: Times New Roman

Page 22: [2] Formatted INSRV 10/1/2010 8:21:00 AM
Font: Times New Roman

Page 22: [2] Formatted INSRV 10/1/2010 8:21:00 AM
Font: Times New Roman

Page 22: [2] Formatted INSRV 10/1/2010 8:21:00 AM
Font: Times New Roman

Page 22: [2] Formatted INSRV 10/1/2010 8:21:00 AM
Font: Times New Roman

Page 22: [2] Formatted INSRV 10/1/2010 8:21:00 AM
Font: Times New Roman

Page 22: [2] Formatted INSRV 10/1/2010 8:21:00 AM
Font: Times New Roman

Page 22: [2] Formatted INSRV 10/1/2010 8:21:00 AM
Font: Times New Roman

Page 22: [2] Formatted INSRV 10/1/2010 8:21:00 AM
Font: Times New Roman, Italic

Page 22: [2] Formatted INSRV 10/1/2010 8:21:00 AM
Font: Times New Roman

Page 22: [2] Formatted INSRV 10/1/2010 8:21:00 AM
Font: Times New Roman

Page 22: [2] Formatted INSRV 10/1/2010 8:21:00 AM
Font: Times New Roman

Page 22: [3] Formatted INSRV 10/1/2010 8:21:00 AM
Font: Times New Roman

Page 22: [3] Formatted INSRV 10/1/2010 8:21:00 AM
Font: Times New Roman

Page 22: [3] Formatted INSRV 10/1/2010 8:21:00 AM
Font: Times New Roman

Page 22: [4] Deleted INSRV 10/1/2010 8:21:00 AM

Page 22: [4] Deleted INSRV 10/1/2010 8:21:00 AM

Page 22: [4] Deleted INSRV 10/1/2010 8:21:00 AM
(20)

Page 22: [5] Formatted INSRV 10/1/2010 8:21:00 AM

Font: Times New Roman

Page 22: [5] Formatted	INSRV	10/1/2010 8:21:00 AM

Font: Times New Roman

Page 22: [5] Formatted	INSRV	10/1/2010 8:21:00 AM

Font: Times New Roman

Page 22: [5] Formatted	INSRV	10/1/2010 8:21:00 AM

Font: Times New Roman

Page 22: [5] Formatted	INSRV	10/1/2010 8:21:00 AM

Font: Times New Roman

Page 22: [6] Formatted	INSRV	10/1/2010 8:22:00 AM

Font: Times New Roman

Page 22: [6] Formatted	INSRV	10/1/2010 8:22:00 AM

Font: Times New Roman

Page 22: [6] Formatted	INSRV	10/1/2010 8:22:00 AM

Font: Times New Roman

Page 22: [6] Formatted	INSRV	10/1/2010 8:22:00 AM

Font: Times New Roman

Page 22: [6] Formatted	INSRV	10/1/2010 8:22:00 AM

Font: Times New Roman

Page 22: [6] Formatted	INSRV	10/1/2010 8:22:00 AM

Font: Times New Roman

Page 22: [6] Formatted	INSRV	10/1/2010 8:22:00 AM

Font: Times New Roman

Page 22: [6] Formatted	INSRV	10/1/2010 8:22:00 AM

Font: Times New Roman

Page 22: [6] Formatted	INSRV	10/1/2010 8:22:00 AM

Font: Times New Roman

Page 22: [6] Formatted	INSRV	10/1/2010 8:22:00 AM

Font: Times New Roman

Page 22: [6] Formatted	INSRV	10/1/2010 8:22:00 AM

Font: Times New Roman

Page 22: [6] Formatted	INSRV	10/1/2010 8:22:00 AM

Font: Times New Roman

Page 22: [6] Formatted	INSRV	10/1/2010 8:22:00 AM

Font: Times New Roman

Page 22: [6] Formatted	INSRV	10/1/2010 8:22:00 AM

Font: Times New Roman

Page 22: [6] Formatted	INSRV	10/1/2010 8:22:00 AM

Font: Times New Roman

Dataset	Population	Intergenic		Introns		Synonymous		Nonsynonymous		HGMD		DM		Probably damaging	
		d_H	Δd	d_H	Δd	d_H	Δd	d_H	Δd	d_H	Δd	d_H	Δd	d_H	Δd
HapMap	**CEU**	**0.03**	**0.013**	**0.03**	**0.022**	**0.08**	**0.043**	**0.09**	*0.016*	**0.09**	*0.052*	*0.07*	*0.031*	**0.08**	*0.055*
	CHB+JPT	**0.03**	**0.013**	**0.03**	**0.02**	**0.08**	**0.052**	**0.11**	*0.015*	**0.09**	*0.026*	**0.11**	*0.062*	**0.14**	**0.086**
	YRI	**0.03**	**0.016**	**0.03**	**0.028**	**0.07**	**0.033**	**0.07**	**0.034**	**0.12**	*0.076*	*0.1*	*0.056*	**0.07**	*0.042*
Resequencing	**AFR**					**0.1**	**0.073**	**0.06**	**0.05**	**0.1**	*0.09*	*0.05*	*0.052*	*0.02*	*0.046*
	CAU					**0.1**	**0.092**	**0.05**	**0.035**	**0.07**	*0.098*	-0.01	*0.039*	*0.04*	*0.026*

Table 1: Summary table for the BGC hallmarks for the HapMap and resequencing SNP datasets. The difference in mean derived allele frequencies between AT→GC and GC→AT SNPs is denoted by d. d_H is the value of d in regions of high recombination. Δd represents the difference in d between the high and low recombination regions. Bold font: values are positive and significantly different from zero, with a p-value < 0.05. Italic font: values are positive but not significantly different from zero. Normal font: values are negative but not significantly different from zero. No cases were found where d_H or Δd were significantly lower than zero.

Figure 1

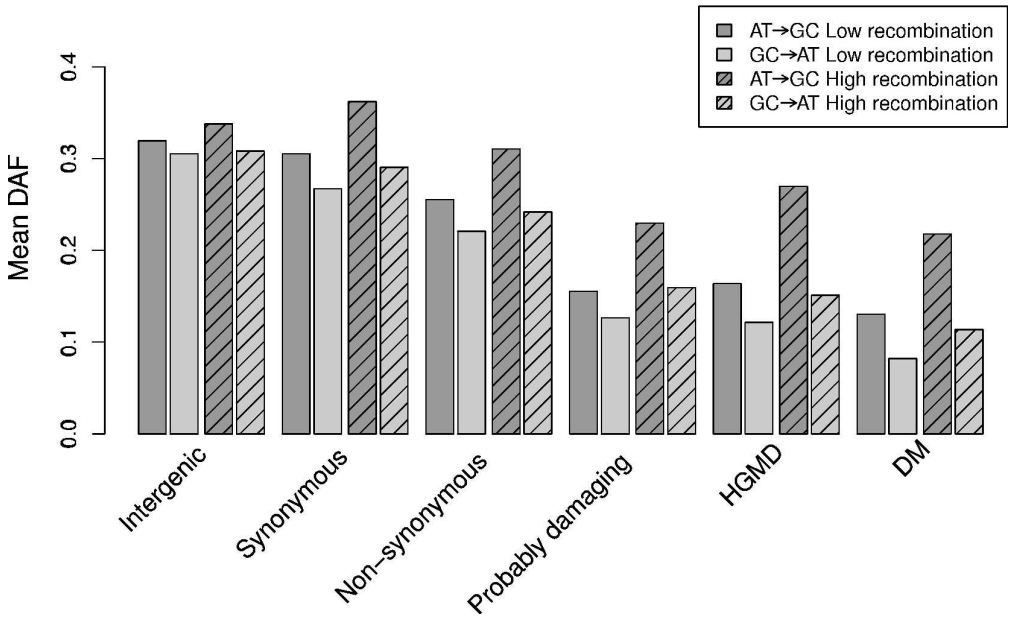

Figure 1. Mean derived allele frequencies for AT→GC and GC→AT alleles in regions of high and low recombination, for the HapMap YRI sample, for different genomic regions and classes of non-synonymous SNPs. Dark gray: AT→GC, light gray: GC→AT mutations. Solid bars: low recombination, hatched bars: high recombination. Probably damaging: HapMap non-synonymous SNPs predicted by Polyphen to be probably damaging. HGMD: entire HGMD dataset. DM: inherited mutations known to be a direct cause of disease (HGMD mutations minus those that have only been associated statistically with disease).
210x144mm (600 x 600 DPI)

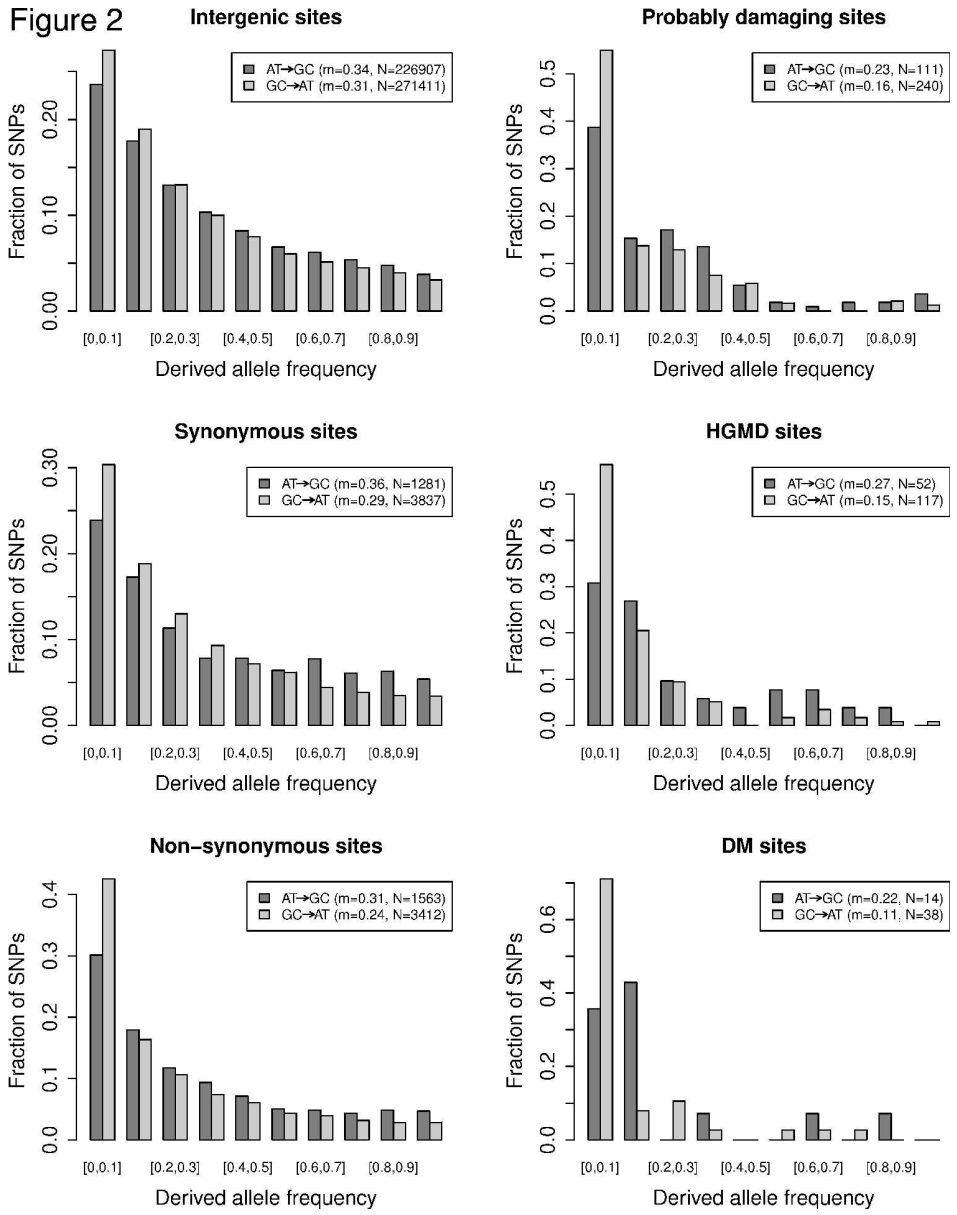

Figure 2. Derived allele frequency spectra for the HapMap YRI sample, for different genomic regions and classes of non-synonymous SNPs. The data presented here relate only to the high recombination class. Dark gray: AT→GC mutations, light gray: GC→AT mutations.
192x245mm (600 x 600 DPI)

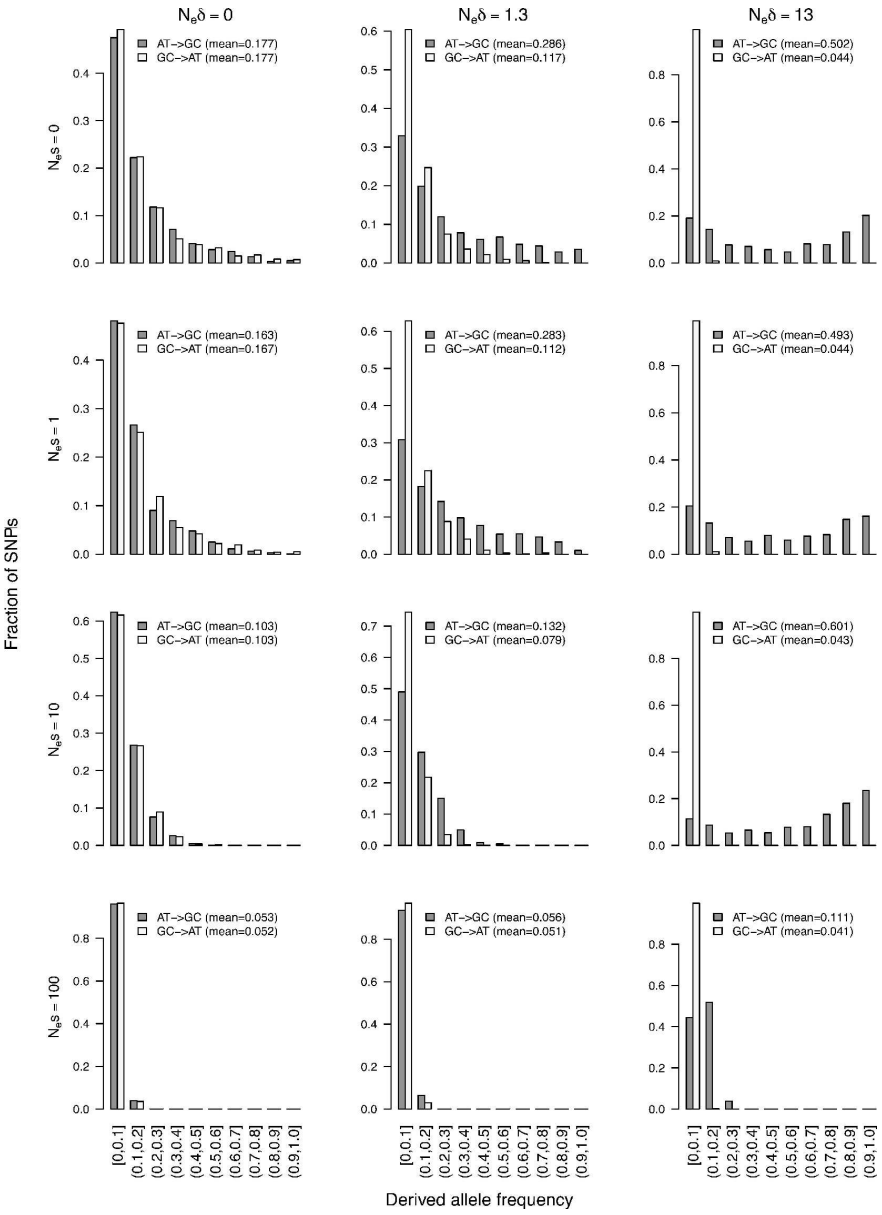

Figure 3. Derived allele frequency spectrum obtained through simulations with different parameter sets. Represented in light gray are the distributions of derived allele frequencies for GC→AT alleles, and in dark gray, those of AT→GC alleles. The population-scaled selection coefficient (Nes) and the population-scaled biased gene conversion parameter (Ne) is indicated for each graph.
210x290mm (600 x 600 DPI)

Necsulea *et al., Human Mutation*

Supplementary Material

Meiotic recombination favors the spreading of deleterious mutations in human populations

Anamaria Necşulea[1], Alexandra Popa[1], David N. Cooper[2], Peter D. Stenson[2], Dominique Mouchiroud[1], Christian Gautier[1], Laurent Duret[1]

[1] Université de Lyon, F-69000, Lyon; Université Lyon 1; CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Évolutive, F-69622, Villeurbanne, France; HELIX, Unité de recherche INRIA.

[2] Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK

Contents

1 Supplementary Text

The effect of sampling bias on the power of gBGC detection

To search for a potential effect of gBGC, we measured the difference between the mean $AT \longrightarrow GC$ and $GC \longrightarrow AT$ frequencies (denoted d) in regions of low or high recombination. We performed this analysis on two independent SNP datasets: the HapMap SNP dataset (Frazer *et al.*, 2007), and the SNP dataset from Lohmueller *et al.* (2008), which was obtained by direct exon resequencing. Both datasets revealed the hallmarks of gBGC (Table 1): i) in regions of high recombination, d is positive; ii) d increases with the recombination rate. This pattern is observed not only for the silent sites but also for the three classes of non-synonymous sites (Table 1).

We noted however that for the two subsets of non-synonymous sites under the strongest purifying selection (DM and probably damaging mutations), the parameter d was lower in the re-sequencing dataset than in the HapMap dataset, and was no longer significantly different from zero (Table 1). We suspected that this difference might be a consequence of the differences between the SNP sampling strategies that were used to prepare the two datasets. On the one hand, the HapMap dataset prioritized using validated SNPs in order to focus resources on common (rather than rare or false positive) candidate SNPs from the public databases (Frazer *et al.*, 2007). This strategy would omit very rare alleles. On the other hand, the SNP dataset from Lohmueller *et al.* (2008) was obtained by direct exon resequencing, and therefore contains no *a priori* bias in allele frequencies. To assess the impact of these two sampling strategies on the detection of biased gene conversion, we performed theoretical simulations.

First we simulated populations (N=10^4) whose genome is subject to mutation, mild negative selection (ω=0.9999) and medium gBGC (δ=0.00013), as described in the main text (Methods). We recorded the allele frequencies at each polymorphic locus (SNP) in these initial populations. Simulations were performed on a large number of loci, so as to obtain a total of more than 10,000 SNPs within each initial population. We sampled N_c chromosomes from these initial populations. Two values were tested for N_c: 120 (60 pairs of chromosomes in a sample, similar to the number of chromosomes in the HapMap samples), 40 (20 pairs of chromosomes in a

sample, similar to the Lohmueller et al. (2008) sample). In order to test the influence of sampling alone, both strategies were compared for the same value of N_c.

To simulate the re-sequencing strategy used by Lohmueller et al. (2008), we randomly drew SNPs in the initial population, and then determined their DAF in the sample of N_c chromosomes. SNPs for which only one allele was present in the sample (*i.e.* that were not detected as polymorphic) were discarded. This random sampling of SNPs was repeated until a dataset of 100 W→S and 100 S→W SNPs was obtained. Then we calculated the statistic d, which is defined as the difference between the mean DAFs of W→S SNPs and of S→W SNPs. To obtain the distribution of the d statistic, this procedure was repeated until at least 60 independent SNP datasets were obtained (Supplementary Table S18).

We used the same procedure to simulate the HapMap strategy, except that rare SNPs (*i.e* SNPs whose DAF in the initial population was lower or higher than given thresholds) were excluded. The threshold values for rare SNPs were chosen in order to have 95% of DAFs in the final sample within the interval [0.03, 0.97].

In the initial simulated population, the derived allele frequency spectrum shows the hallmark of gBGC: DAFs are significantly higher for W→S SNPs than for S→W SNPs (Figure 3). This excess of high-frequency W→S SNPs (*i.e.* a positive d statistic) was detected with both sampling strategies. However, independently of the number of sampled chromosomes (N_c=120 or N_c=40), the d statistic obtained by the re-sequencing strategy was on average lower than that obtained with the HapMap strategy (Supplementary Figure S1. For $N_c = 40$ - Resequencing: d=0.1129; HapMap: d=0.146, Wilcoxon rank sum test: p=$1e^{-4}$. For $N_c = 120$ - Resequencing: d=0.0893; HapMap: d=0.1318, Wilcoxon rank sum test: p=0.). This indicates that all else being equal, the HapMap SNP sampling strategy has a greater power to detect the effect of gBGC than the re-sequencing strategy. This can be explained by the fact that the impact of gBGC is more readily detectable on the upper part of the DAF spectrum.

2 Supplementary Figures

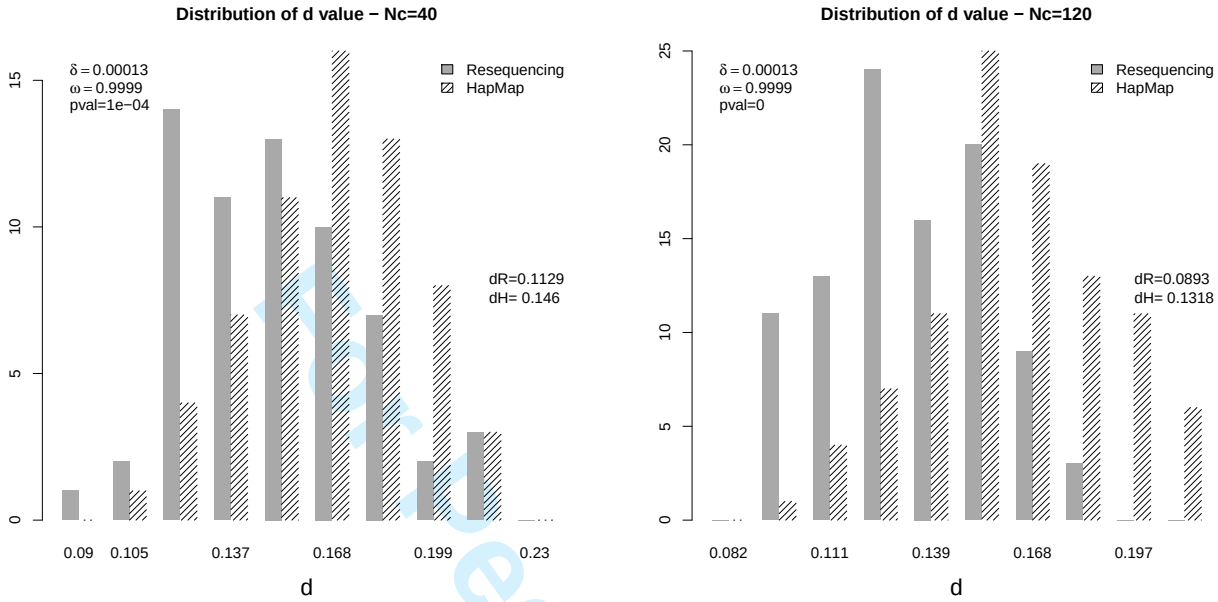

Figure S1: The distribution of the statistics d in the case of an initial spectrum frequencies obtained with values of $\delta = 0.00013$ and $\omega = 0.9999$. Left panel: simulations for $N_c = 40$; right panel: for $N_c = 120$. Gray bars: the histogram of d in the case of HapMap sampling strategy. Hatched bars: the histogram of resequencing sampling strategy. The p-value corresponds to the Wilcoxon rank sum test, for a comparison of medians of the two distributions. The mean value of d is also given in the case of HapMap (dH) and resequencing (dR) sampling strategies.

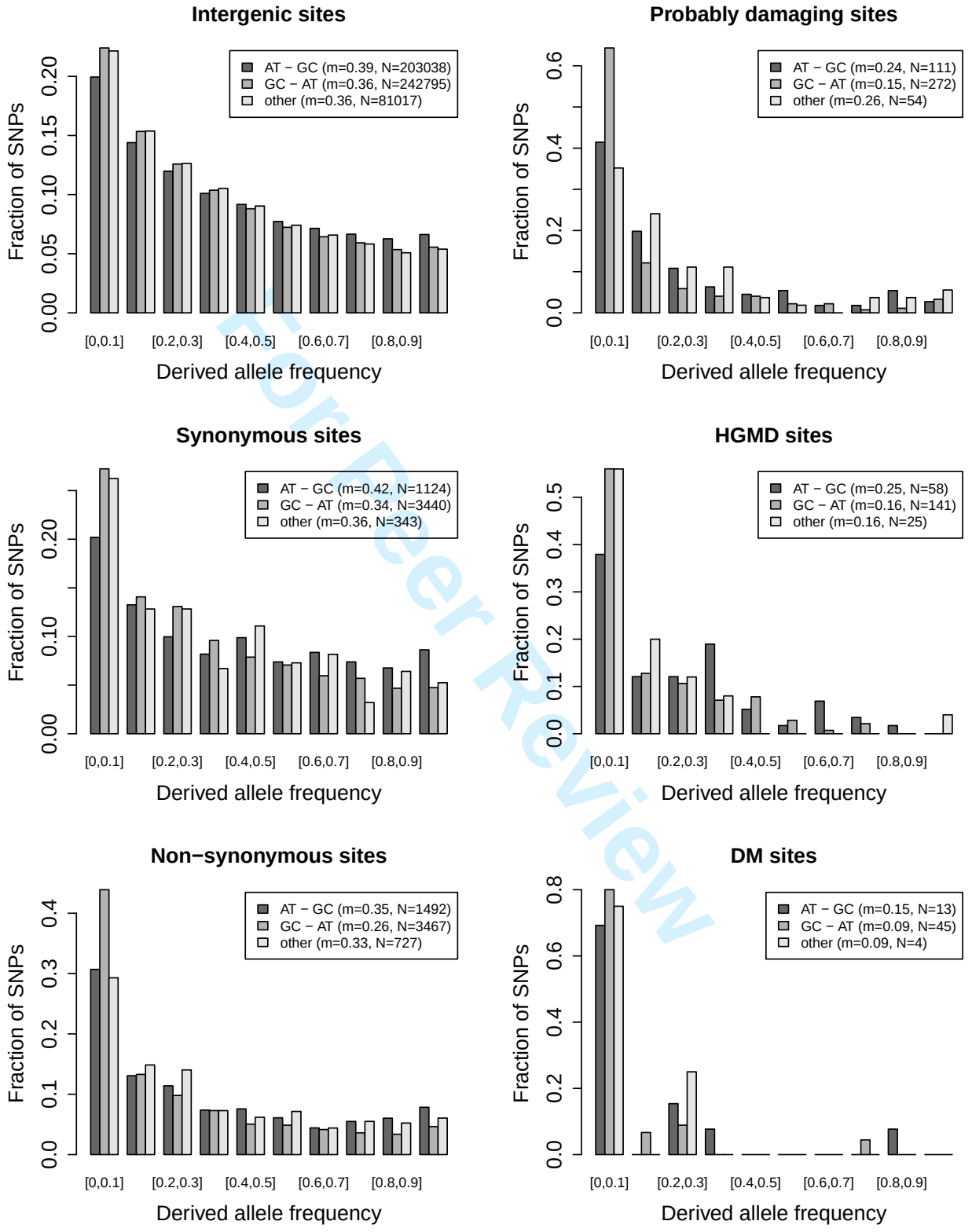

Figure S2: Derived allele frequencies spectra obtained for the HapMap CEU sample, for high recombination regions (as defined by the distance to recombination hotspots). The ancestral and derived alleles were determined using a maximum likelihood method that takes into account CpG hypermutability (Duret and Arndt, 2008). m represents the mean derived allele frequency, and N is the number of SNPs in each category.

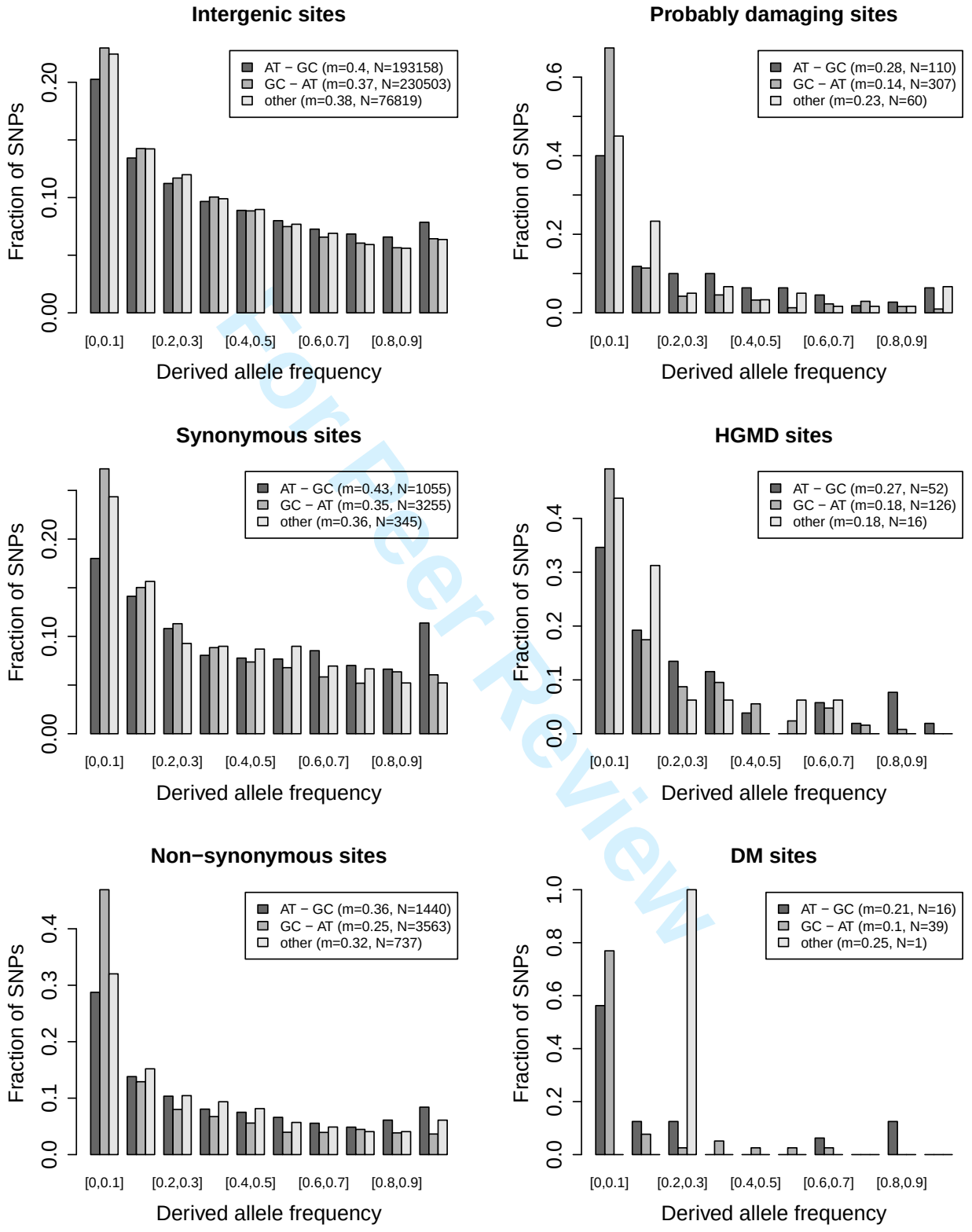

Figure S3: Derived allele frequencies spectra obtained for the HapMap CHB+JPT sample, for high recombination regions (as defined by the distance to recombination hotspots). The ancestral and derived alleles were determined using a maximum likelihood method that takes into account CpG hypermutability (Duret and Arndt, 2008). m represents the mean derived allele frequency, and N is the number of SNPs in each category.

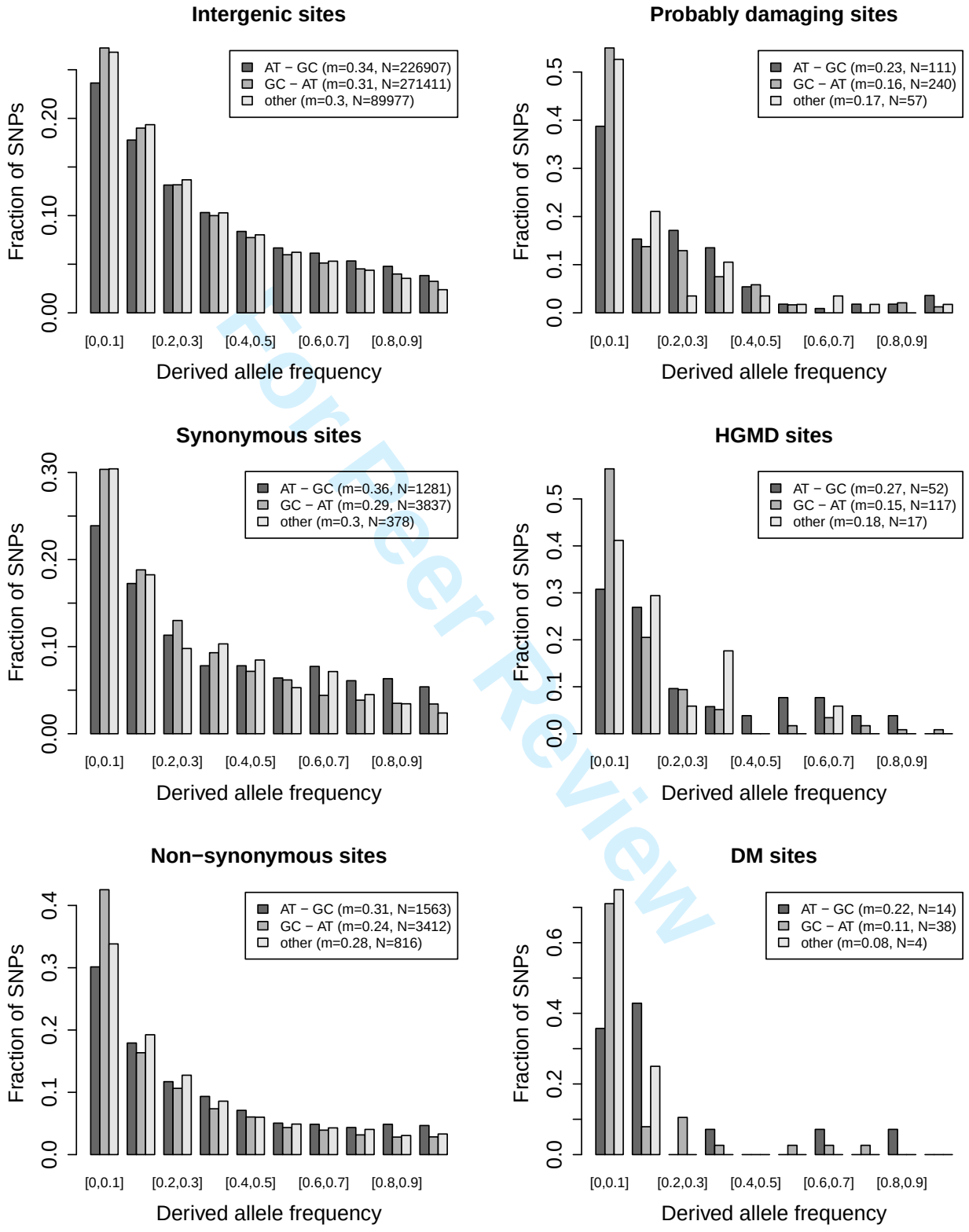

Figure S4: Derived allele frequencies spectra obtained for the HapMap YRI sample, for high recombination regions (as defined by the distance to recombination hotspots). The ancestral and derived alleles were determined using a maximum likelihood method that takes into account CpG hypermutability (Duret and Arndt, 2008). m represents the mean derived allele frequency, and N is the number of SNPs in each category.

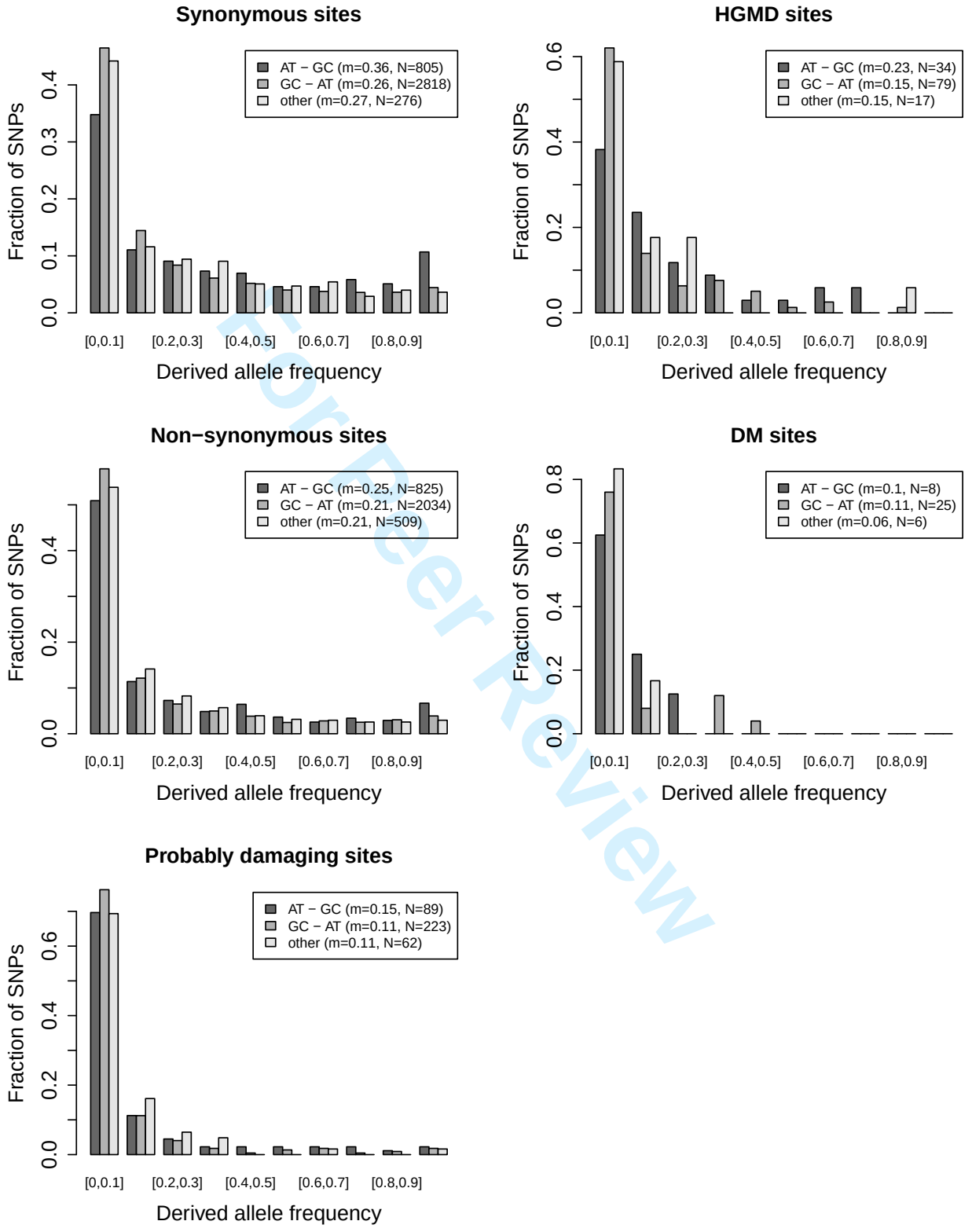

Figure S5: Derived allele frequencies spectra obtained for the Lohmueller *et al.* (2008) CAU sample, for high recombination regions (as defined by the distance to recombination hotspots). The ancestral and derived alleles were determined using a maximum likelihood method that takes into account CpG hypermutability (Duret and Arndt, 2008). m represents the mean derived allele frequency, and N is the number of SNPs in each category.

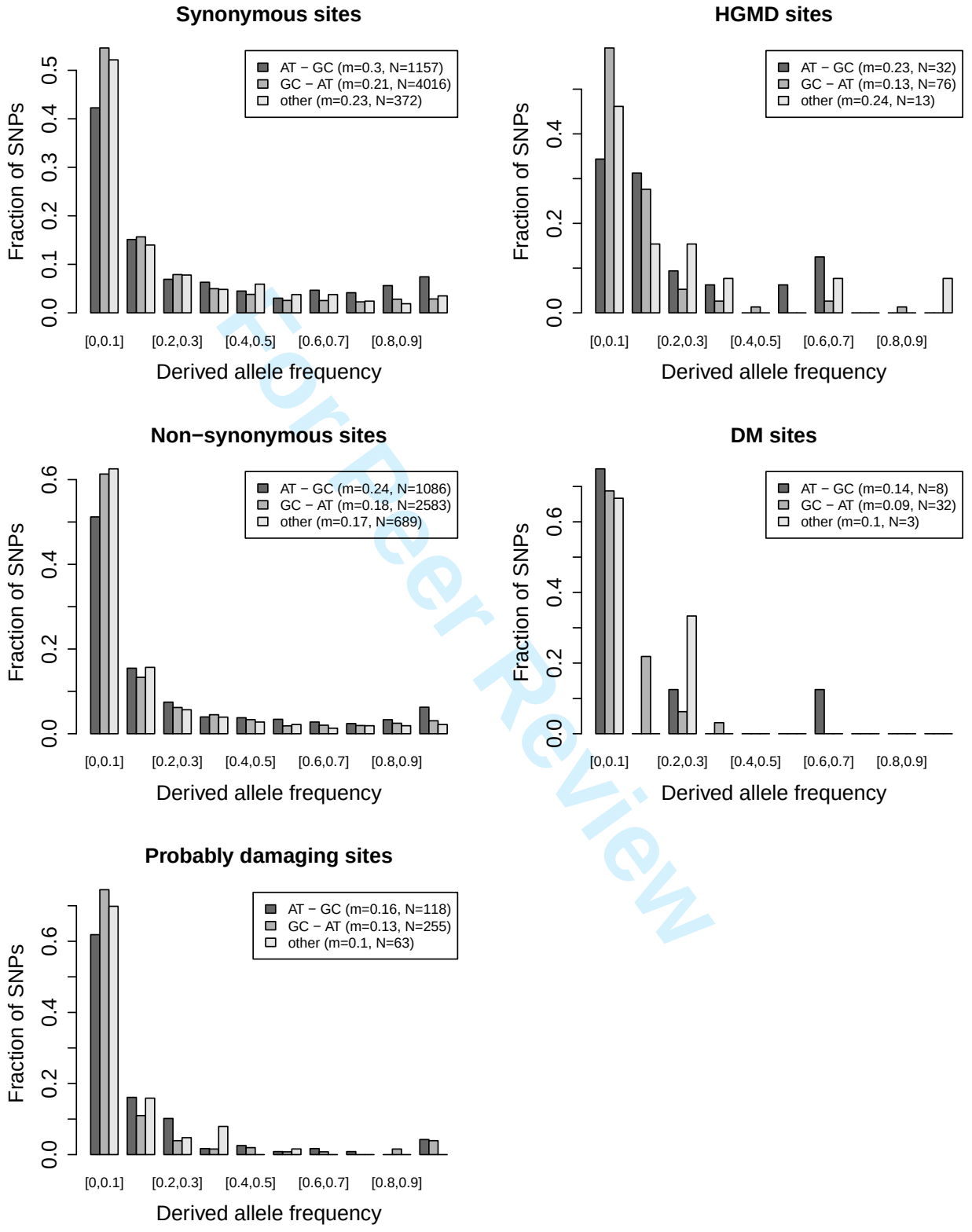

Figure S6: Derived allele frequencies spectra obtained for the Lohmueller *et al.* (2008) AFR sample, for high recombination regions (as defined by the distance to recombination hotspots). The ancestral and derived alleles were determined using a maximum likelihood method that takes into account CpG hypermutability (Duret and Arndt, 2008). m represents the mean derived allele frequency, and N is the number of SNPs in each category.

3 Supplementary Tables

Type	CEU	CHB+JPT	YRI	All samples
Intergenic	1,705,876	1,621,044	1,903,673	2,151,095
Introns	1,044,507	998,219	1,168,601	1,329,002
Synonymous	15,313	14,545	17,155	20,095
Non-synonymous	18,251	18,467	18,594	24,609
Other	31,011	29,405	34,093	41,576
Total	2,814,958	2,681,680	3,142,116	3,566,377

Table S1: SNP dataset from HapMap release 27. 5' and 3' UTR exons are excluded from our dataset. Note that the total sample size is given here and that further restrictions are applied when computing the DAF spectrum: constant 5' - 3' context (*i.e.* positions with two neighboring SNPs were removed, and we required that the human and chimpanzee nucleotides should be identical) and non-ambiguous ancestral allele prediction.

Type	AFR	CAU
Synonymous	17,011	11,931
Non-synonymous	13,707	10,583
Total	30,718	22,514

Table S2: SNP dataset from Lohmueller *et al.*, 2008. Note that the total sample size is given here and that further restrictions are applied when computing the DAF spectrum: constant 5' - 3' context, (*i.e.* positions with two neighboring SNPs were removed, and we required that the human and chimpanzee nucleotides should be identical) and non-ambiguous ancestral allele prediction.

Type	All HGMD	CEU	CHB+JPT	YRI	AFR	CAU
DM	41,949 (33,096)	225 (221)	213 (207)	215 (210)	163 (163)	142 (142)
DP	837 (801)	492 (477)	441 (425)	412 (399)	241 (241)	274 (274)
FP	923 (866)	168 (162)	140 (135)	150 (145)	71 (71)	73 (73)
DFP	51 (51)	36 (36)	32 (32)	31 (31)	23 (23)	22 (22)
Total	43,760 (34,814)	921 (896)	826 (799)	808 (785)	498 (498)	511 (511)

Table S3: Number of non-synonymous disease-associated mutations in HGMD and found within our SNP datasets. The numbers in parantheses represent the number of missense mutations (the remaining ones are nonsense mutations). Note that this table includes mutations for which the disease-associated allele is ancestral, although in the derived allele frequencies spectra we include only positions for which the disease-associated allele is derived.

Type	CEU	CHB+JPT	YRI	AFR	CAU
Benign	12,338	12,468	12,681	9,698	7,366
Possibly damaging	2,566	2,688	2,597	2,366	1,829
Probably damaging	1,591	1,721	1,518	1,361	1,168
Total	16,495	16,877	16,796	13,425	10,363

Table S4: PolyPhen predictions for the non-synonymous SNPs in our dataset.

Type	CEU	CHB+JPT	YRI	AFR	CAU
Intergenic	56	88	60		
Intron	56	78	52		
Synonymous	90	84	92	18	18
Non-synonymous	92	92	90	18	18
HGMD	100	144	100	18	18
DM	100	148	106	18	18
Probably damaging	98	148	94	18	18

Table S5: The minimum number of genotyped chromosomes for each SNP dataset.

Dataset	Population	Intergenic		Introns		Synonymous		Non-synonymous		HGMD		DM		Probably damaging	
		d_H	Δd	d_H	Δd	d_H	Δd	d_H	Δd	d_H	Δd	d_H	Δd	d_H	Δd
Maximum likelihood method (Duret and Arndt, 2008)															
HapMap	CEU	0.03	0.013	0.03	0.022	0.08	0.043	0.09	0.016	0.09	0.052	0.07	0.031	0.08	0.055
	CHB+JPT	0.03	0.013	0.03	0.02	0.08	0.052	0.11	0.015	0.09	0.026	0.11	0.062	0.14	0.086
	YRI	0.03	0.016	0.03	0.028	0.07	0.033	0.07	0.034	0.12	0.076	0.1	0.056	0.07	0.042
Resequencing	AFR					0.1	0.073	0.06	0.05	0.1	0.09	0.05	0.052	0.02	0.046
	CAU					0.1	0.092	0.05	0.035	0.07	0.098	-0.01	0.039	0.04	0.026
SFS correction with context-dependent model (Hernandez *et al.*, 2007)															
HapMap	CEU	0.04	0.012	0.05	0.019	0.06	0.02	0.07	0.005	0.07	0.034	0.02	-0.016	0.08	0.038
	CHB+JPT	0.05	0.013	0.05	0.018	0.07	0.034	0.09	0.003	0.08	0.034	0.08	0.041	0.13	0.06
	YRI	0.05	0.014	0.06	0.023	0.05	0.011	0.05	0.022	0.1	0.071	0.04	0.006	0.07	0.022
Resequencing	AFR					0.06	0.048	0.04	0.027	0.09	0.074	0.05	0.047	0.04	0.026
	CAU					0.06	0.061	0.03	0.011	0.06	0.06	-0.02	0.042	0.06	0.031

Table S6: Summary table for the BGC hallmarks. d_H represents the difference in mean derived allele frequencies between AT→GC and GC→AT SNPs, for high recombination regions. Δd represents the difference in d between high and low recombination regions. The recombination classes were computed as a function of the distance to recombination hotspots. Dark green: d and Δd are positive, with a p-value < 0.05. Light green: d and Δd are positive, with a p-value ≥ 0.05. Light red: d and Δd are negative, with a p-value $>= 0.05$. No cases were found where d and Δd are negative, with a p-value < 0.05.

Dataset	Population	Intergenic		Introns		Synonymous		Non-synonymous		HGMD		DM		Probably damaging	
		d_H	Δd	d_H	Δd	d_H	Δd	d_H	Δd	d_H	Δd	d_H	Δd	d_H	Δd
Maximum likelihood method (Duret and Arndt, 2008)															
HapMap	CEU	0.02	0.012	0.03	0.018	0.05	0.032	0.09	0.017	0.12	0.07	0.03	-0.038	0.04	0.058
	CHB+JPT	0.03	0.012	0.02	0.016	0.05	0.031	0.11	0.009	0.14	0.067	0.12	0.001	0.08	0.074
	YRI	0.02	0.013	0.02	0.024	0.03	0.013	0.05	0.016	0.1	0.056	-0.04	-0.132	0.03	0.019
Resequencing	AFR					0.05	0.059	0.03	0.023	0.11	0.043	0.1	-0.037	0.05	0.036
	CAU					0.06	0.067	0.04	0.047	0.12	0.102	0.01	-0.002	0.05	0.051
SFS correction with context-dependent model (Hernandez *et al.*, 2007)															
HapMap	CEU	0.03	0.011	0.03	0.015	0.04	0.013	0.08	0.015	0.13	0.057	0.03	-0.041	0.03	0.04
	CHB+JPT	0.03	0.012	0.03	0.014	0.04	0.017	0.09	0	0.13	0.079	0.13	0.012	0.06	0.045
	YRI	0.04	0.012	0.04	0.019	0.02	-0.008	0.05	0.014	0.08	0.043	-0.06	-0.149	0.04	0.04
Resequencing	AFR					0.04	0.042	0.04	0.013	0.11	0.035	0.1	-0.036	0.07	0.047
	CAU					0.04	0.049	0.04	0.033	0.11	0.047	0.02	-0.001	0.06	0.052

Table S7: Summary table for the BGC hallmarks. d_H represents the difference in mean derived allele frequencies between AT→GC and GC→AT SNPs, for high recombination regions. Δd represents the difference in d between high and low recombination regions. The recombination classes were computed as a function of the distance to recombination hotspots. Only SNPs found within a non-CpG context (*i.e.* the 5' neighbor is different from C and the 3' neighbor is different from G) were considered here. Dark green: d and Δd are positive, with a p-value < 0.05. Light green: d and Δd are positive, with a p-value ≥ 0.05. Light red: d and Δd are negative, with a p-value $>= 0.05$. No cases were found where d and Δd are negative, with a p-value < 0.05.

Region	Nb. $AT \rightarrow GC$	Mean $AT \rightarrow GC$	Nb. $GC \rightarrow AT$	Mean $GC \rightarrow AT$	d_H	P-value
Maximum likelihood method (Duret and Arndt, 2008)						
Intergenic	203038	0.39	242795	0.36	0.03	0 / 1000
Intron	127478	0.39	150521	0.36	0.03	0 / 1000
Synonymous	1124	0.42	3440	0.34	0.08	0 / 1000
Non-synonymous	1492	0.35	3467	0.26	0.09	0 / 1000
PolyPhen probably damaging	111	0.24	272	0.15	0.08	0 / 1000
HGMD	58	0.25	141	0.16	0.09	6 / 1000
DM	13	0.15	45	0.09	0.07	105 / 1000
SFS correction with context-dependent model (Hernandez *et al.*, 2007)						
Intergenic	214565	0.39	217374	0.35	0.04	0 / 1000
Intron	134207	0.39	134843	0.34	0.05	0 / 1000
Synonymous	1113	0.39	3245	0.34	0.06	0 / 1000
Non-synonymous	1443	0.33	3061	0.27	0.07	0 / 1000
PolyPhen probably damaging	108	0.24	232	0.16	0.08	5 / 1000
HGMD	59	0.25	130	0.17	0.07	8 / 1000
DM	12	0.11	39	0.09	0.02	354 / 1000

Table S8: Derived allele frequencies for $AT \rightarrow GC$ and $GC \rightarrow AT$ SNPs, for the CEU sample from HapMap, for high recombination regions. The recombination classes were defined based on the distance to recombination hotspots.

Region	Nb. $AT \rightarrow GC$	Mean $AT \rightarrow GC$	Nb. $GC \rightarrow AT$	Mean $GC \rightarrow AT$	d_H	P-value
Maximum likelihood method (Duret and Arndt, 2008)						
Intergenic	193158	0.4	230503	0.37	0.03	0 / 1000
Intron	121829	0.4	143749	0.37	0.03	0 / 1000
Synonymous	1055	0.43	3255	0.35	0.08	0 / 1000
Non-synonymous	1440	0.36	3563	0.25	0.11	0 / 1000
PolyPhen probably damaging	110	0.28	307	0.14	0.14	0 / 1000
HGMD	52	0.27	126	0.18	0.09	9 / 1000
DM	16	0.21	39	0.1	0.11	22 / 1000
SFS correction with context-dependent model (Hernandez *et al.*, 2007)						
Intergenic	205085	0.4	207474	0.36	0.05	0 / 1000
Intron	129054	0.4	129683	0.35	0.05	0 / 1000
Synonymous	1054	0.41	3121	0.35	0.07	0 / 1000
Non-synonymous	1406	0.34	3204	0.25	0.09	0 / 1000
PolyPhen probably damaging	105	0.28	264	0.15	0.13	0 / 1000
HGMD	52	0.27	115	0.19	0.08	13 / 1000
DM	14	0.18	36	0.1	0.08	86 / 1000

Table S9: Derived allele frequencies for $AT \rightarrow GC$ and $GC \rightarrow AT$ SNPs, for the CHB+JPT sample from HapMap, for high recombination regions. The recombination classes were defined based on the distance to recombination hotspots.

Region	Nb. $AT \rightarrow GC$	Mean $AT \rightarrow GC$	Nb. $GC \rightarrow AT$	Mean $GC \rightarrow AT$	d_H	P-value
Maximum likelihood method (Duret and Arndt, 2008)						
Intergenic	226907	0.34	271411	0.31	0.03	0 / 1000
Intron	142683	0.34	168833	0.3	0.03	0 / 1000
Synonymous	1281	0.36	3837	0.29	0.07	0 / 1000
Non-synonymous	1563	0.31	3412	0.24	0.07	0 / 1000
PolyPhen probably damaging	111	0.23	240	0.16	0.07	1 / 1000
HGMD	52	0.27	117	0.15	0.12	0 / 1000
DM	14	0.22	38	0.11	0.1	65 / 1000
SFS correction with context-dependent model (Hernandez *et al.*, 2007)						
Intergenic	240916	0.33	244249	0.28	0.05	0 / 1000
Intron	150974	0.33	151760	0.27	0.06	0 / 1000
Synonymous	1298	0.33	3648	0.28	0.05	0 / 1000
Non-synonymous	1558	0.29	3063	0.24	0.05	0 / 1000
PolyPhen probably damaging	113	0.23	202	0.16	0.07	3 / 1000
HGMD	55	0.27	104	0.16	0.1	2 / 1000
DM	14	0.16	32	0.12	0.04	358 / 1000

Table S10: Derived allele frequencies for $AT \rightarrow GC$ and $GC \rightarrow AT$ SNPs, for the YRI sample from HapMap, for high recombination regions. The recombination classes were defined based on the distance to recombination hotspots.

Region	Nb. $AT \rightarrow GC$	Mean $AT \rightarrow GC$	Nb. $GC \rightarrow AT$	Mean $GC \rightarrow AT$	d_H	P-value
Maximum likelihood method (Duret and Arndt, 2008)						
Synonymous	1157	0.3	4016	0.21	0.1	0 / 1000
Non-synonymous	1086	0.24	2583	0.18	0.06	0 / 1000
PolyPhen probably damaging	118	0.16	255	0.13	0.02	159 / 1000
HGMD	32	0.23	76	0.13	0.1	9 / 1000
DM	8	0.14	32	0.09	0.05	100 / 1000
SFS correction with context-dependent model (Hernandez *et al.*, 2007)						
Synonymous	1026	0.28	3363	0.22	0.06	0 / 1000
Non-synonymous	955	0.23	2072	0.19	0.04	0 / 1000
PolyPhen probably damaging	104	0.16	187	0.12	0.04	17 / 1000
HGMD	30	0.25	65	0.15	0.09	13 / 1000
DM	8	0.16	26	0.12	0.05	266 / 1000

Table S11: Derived allele frequencies for $AT \rightarrow GC$ and $GC \rightarrow AT$ SNPs, for the AFR sample from Lohmueller *et al.*, 2008, for high recombination regions. The recombination classes were defined based on the distance to recombination hotspots.

Region	Nb. $AT \rightarrow GC$	Mean $AT \rightarrow GC$	Nb. $GC \rightarrow AT$	Mean $GC \rightarrow AT$	d_H	P-value
Maximum likelihood method (Duret and Arndt, 2008)						
Synonymous	805	0.36	2818	0.26	0.1	0 / 1000
Non-synonymous	825	0.25	2034	0.21	0.05	0 / 1000
PolyPhen probably damaging	89	0.15	223	0.11	0.04	69 / 1000
HGMD	34	0.23	79	0.15	0.07	33 / 1000
DM	8	0.1	25	0.11	-0.01	635 / 1000
SFS correction with context-dependent model (Hernandez *et al.*, 2007)						
Synonymous	675	0.35	2252	0.29	0.06	0 / 1000
Non-synonymous	662	0.27	1493	0.24	0.03	0 / 1000
PolyPhen probably damaging	66	0.2	149	0.14	0.06	6 / 1000
HGMD	31	0.26	62	0.2	0.06	96 / 1000
DM	8	0.11	17	0.13	-0.02	797 / 1000

Table S12: Derived allele frequencies for $AT \rightarrow GC$ and $GC \rightarrow AT$ SNPs, for the CAU sample from Lohmueller *et al.*, 2008, for high recombination regions. The recombination classes were defined based on the distance to recombination hotspots.

Region	Low recombination			High recombination			Δd	P-value
	$AT \rightarrow GC$	$GC \rightarrow AT$	d	$AT \rightarrow GC$	$GC \rightarrow AT$	d		
Maximum likelihood method (Duret and Arndt, 2008)								
Intergenic	0.38	0.36	0.01	0.39	0.36	0.03	0.013	0 / 1000
Intron	0.37	0.36	0.01	0.39	0.36	0.03	0.022	0 / 1000
Synonymous	0.37	0.33	0.04	0.42	0.34	0.08	0.043	1 / 1000
Non-synonymous	0.3	0.22	0.07	0.35	0.26	0.09	0.016	122 / 1000
PolyPhen probably damaging	0.16	0.13	0.03	0.24	0.15	0.08	0.055	95 / 1000
HGMD	0.16	0.13	0.03	0.25	0.16	0.09	0.052	120 / 1000
DM	0.09	0.05	0.04	0.15	0.09	0.07	0.031	347 / 1000
SFS correction with context-dependent model (Hernandez *et al.*, 2007)								
Intergenic	0.38	0.34	0.03	0.39	0.35	0.04	0.012	0 / 1000
Intron	0.36	0.34	0.03	0.39	0.34	0.05	0.019	0 / 1000
Synonymous	0.36	0.32	0.04	0.39	0.34	0.06	0.02	80 / 1000
Non-synonymous	0.29	0.23	0.06	0.33	0.27	0.07	0.005	358 / 1000
PolyPhen probably damaging	0.17	0.13	0.04	0.24	0.16	0.08	0.038	163 / 1000
HGMD	0.19	0.15	0.04	0.25	0.17	0.07	0.034	191 / 1000
DM	0.1	0.06	0.03	0.11	0.09	0.02	-0.016	609 / 1000

Table S13: Derived allele frequencies for $AT \rightarrow GC$ and $GC \rightarrow AT$ SNPs, for the CEU sample from HapMap, for regions of low and high recombination. The regions of low and high recombination were defined based on the distance from SNPs to recombination hotspots.

Region	Low recombination			High recombination			Δd	P-value
	$AT \rightarrow GC$	$GC \rightarrow AT$	d	$AT \rightarrow GC$	$GC \rightarrow AT$	d		
Maximum likelihood method (Duret and Arndt, 2008)								
Intergenic	0.39	0.37	0.02	0.4	0.37	0.03	0.013	0 / 1000
Intron	0.38	0.37	0.01	0.4	0.37	0.03	0.02	0 / 1000
Synonymous	0.37	0.34	0.03	0.43	0.35	0.08	0.052	0 / 1000
Non-synonymous	0.31	0.22	0.1	0.36	0.25	0.11	0.015	145 / 1000
PolyPhen probably damaging	0.17	0.12	0.06	0.28	0.14	0.14	0.086	23 / 1000
HGMD	0.22	0.15	0.06	0.27	0.18	0.09	0.026	319 / 1000
DM	0.1	0.05	0.05	0.21	0.1	0.11	0.062	319 / 1000
SFS correction with context-dependent model (Hernandez *et al.*, 2007)								
Intergenic	0.39	0.35	0.03	0.4	0.36	0.05	0.013	0 / 1000
Intron	0.38	0.35	0.03	0.4	0.35	0.05	0.018	0 / 1000
Synonymous	0.36	0.33	0.03	0.41	0.35	0.07	0.034	5 / 1000
Non-synonymous	0.31	0.22	0.09	0.34	0.25	0.09	0.003	420 / 1000
PolyPhen probably damaging	0.19	0.12	0.07	0.28	0.15	0.13	0.06	80 / 1000
HGMD	0.22	0.17	0.05	0.27	0.19	0.08	0.034	287 / 1000
DM	0.1	0.06	0.04	0.18	0.1	0.08	0.041	226 / 1000

Table S14: Derived allele frequencies for $AT \rightarrow GC$ and $GC \rightarrow AT$ SNPs, for the CHB+JPT sample from HapMap, for regions of low and high recombination. The regions of low and high recombination were defined based on the distance from SNPs to recombination hotspots.

Region	Low recombination			High recombination			Δd	P-value
	$AT \rightarrow GC$	$GC \rightarrow AT$	d	$AT \rightarrow GC$	$GC \rightarrow AT$	d		
Maximum likelihood method (Duret and Arndt, 2008)								
Intergenic	0.32	0.31	0.01	0.34	0.31	0.03	0.016	0 / 1000
Intron	0.3	0.3	0	0.34	0.3	0.03	0.028	0 / 1000
Synonymous	0.31	0.27	0.04	0.36	0.29	0.07	0.033	4 / 1000
Non-synonymous	0.26	0.22	0.03	0.31	0.24	0.07	0.034	1 / 1000
PolyPhen probably damaging	0.16	0.13	0.03	0.23	0.16	0.07	0.042	142 / 1000
HGMD	0.16	0.12	0.04	0.27	0.15	0.12	0.076	74 / 1000
DM	0.13	0.08	0.05	0.22	0.11	0.1	0.056	259 / 1000
SFS correction with context-dependent model (Hernandez *et al.*, 2007)								
Intergenic	0.31	0.27	0.04	0.33	0.28	0.05	0.014	0 / 1000
Intron	0.3	0.26	0.03	0.33	0.27	0.06	0.023	0 / 1000
Synonymous	0.29	0.25	0.04	0.33	0.28	0.05	0.011	190 / 1000
Non-synonymous	0.25	0.22	0.03	0.29	0.24	0.05	0.022	28 / 1000
PolyPhen probably damaging	0.18	0.13	0.04	0.23	0.16	0.07	0.022	250 / 1000
HGMD	0.18	0.14	0.03	0.27	0.16	0.1	0.071	70 / 1000
DM	0.14	0.1	0.03	0.16	0.12	0.04	0.006	437 / 1000

Table S15: Derived allele frequencies for $AT \rightarrow GC$ and $GC \rightarrow AT$ SNPs, for the YRI sample from HapMap, for regions of low and high recombination. The regions of low and high recombination were defined based on the distance from SNPs to recombination hotspots.

Region	Low recombination			High recombination			Δd	P-value
	$AT \rightarrow GC$	$GC \rightarrow AT$	d	$AT \rightarrow GC$	$GC \rightarrow AT$	d		
Maximum likelihood method (Duret and Arndt, 2008)								
Synonymous	0.24	0.21	0.02	0.3	0.21	0.1	0.073	0 / 1000
Non-synonymous	0.18	0.17	0.01	0.24	0.18	0.06	0.05	0 / 1000
PolyPhen probably damaging	0.1	0.12	-0.02	0.16	0.13	0.02	0.046	64 / 1000
HGMD	0.16	0.15	0.01	0.23	0.13	0.1	0.09	77 / 1000
DM	0.16	0.16	0	0.14	0.09	0.05	0.052	401 / 1000
SFS correction with context-dependent model (Hernandez *et al.*, 2007)								
Synonymous	0.23	0.21	0.02	0.28	0.22	0.06	0.048	0 / 1000
Non-synonymous	0.19	0.17	0.01	0.23	0.19	0.04	0.027	10 / 1000
PolyPhen probably damaging	0.12	0.11	0.01	0.16	0.12	0.04	0.026	152 / 1000
HGMD	0.18	0.16	0.02	0.25	0.15	0.09	0.074	105 / 1000
DM	0.19	0.19	0	0.16	0.12	0.05	0.047	414 / 1000

Table S16: Derived allele frequencies for $AT \rightarrow GC$ and $GC \rightarrow AT$ SNPs, for the AFR sample from Lohmueller *et al.*, 2008, for regions of low and high recombination. The regions of low and high recombination were defined based on the distance from SNPs to recombination hotspots.

Region	Low recombination			High recombination			Δd	P-value
	$AT \rightarrow GC$	$GC \rightarrow AT$	d	$AT \rightarrow GC$	$GC \rightarrow AT$	d		
Maximum likelihood method (Duret and Arndt, 2008)								
Synonymous	0.28	0.27	0.01	0.36	0.26	0.1	0.092	0 / 1000
Non-synonymous	0.2	0.19	0.01	0.25	0.21	0.05	0.035	12 / 1000
PolyPhen probably damaging	0.1	0.09	0.01	0.15	0.11	0.04	0.026	229 / 1000
HGMD	0.12	0.15	-0.02	0.23	0.15	0.07	0.098	57 / 1000
DM	0.06	0.1	-0.05	0.1	0.11	-0.01	0.039	294 / 1000
SFS correction with context-dependent model (Hernandez *et al.*, 2007)								
Synonymous	0.29	0.29	0	0.35	0.29	0.06	0.061	0 / 1000
Non-synonymous	0.24	0.21	0.02	0.27	0.24	0.03	0.011	253 / 1000
PolyPhen probably damaging	0.15	0.12	0.03	0.2	0.14	0.06	0.031	193 / 1000
HGMD	0.19	0.19	0	0.26	0.2	0.06	0.06	151 / 1000
DM	0.08	0.14	-0.06	0.11	0.13	-0.02	0.042	256 / 1000

Table S17: Derived allele frequencies for $AT \rightarrow GC$ and $GC \rightarrow AT$ SNPs, for the CAU sample from Lohmueller *et al.*, 2008, for regions of low and high recombination. The regions of low and high recombination were defined based on the distance from SNPs to recombination hotspots.

Initial frequencies	$\delta = 0$ $\omega = 1$		$\delta = 0.00013$ $\omega = 0.9999$	
No of chromosomes	$N_c = 40$	$N_c = 120$	$N_c = 40$	$N_c = 120$
No of classes	71	92	64	97
Range initial freq.	0.061-0.939	0.036-0.964	0.055-0.945	0.031-0.969

Table S18: Table of the number of classes each of size 100 polymorphic loci and the range of initial frequencies for the HapMap biased sampling strategy. These variables depend on the values of gBGC (δ) and the fitness of the derived allele (ω) which were used to obtain the initial distribution of derived alleles, as well as the number of chromosomes analyzed.

HapMap sample	Q→H		H→Q		E→D		D→E		F→L	
	AT→GC	GC→AT	AT→GC	GC→AT	AT→GC	GC→AT	AT→GC	GC→AT	AT→GC	GC→AT
CEU	0.19 (18)	0.16 (66)	0.35 (12)	0.23 (31)	0.35 (32)	0.24 (71)	0.22 (22)	0.19 (46)	0.35 (88)	0.32 (23)
CHB+JPT	0.18 (12)	0.12 (81)	0.28 (10)	0.31 (25)	0.25 (37)	0.23 (81)	0.23 (18)	0.21 (53)	0.35 (86)	0.2 (31)
YRI	0.22 (16)	0.14 (60)	0.26 (16)	0.23 (29)	0.23 (35)	0.24 (75)	0.16 (23)	0.24 (48)	0.31 (102)	0.29 (19)

HapMap sample	L→F		S→R		R→S		N→K		K→N	
	AT→GC	GC→AT	AT→GC	GC→AT	AT→GC	GC→AT	AT→GC	GC→AT	AT→GC	GC→AT
CEU	0.2 (7)	0.23 (197)	0.34 (29)	0.31 (23)	0.29 (8)	0.24 (42)	0.42 (15)	0.2 (34)	0.22 (29)	0.18 (27)
CHB+JPT	0.3 (7)	0.2 (229)	0.29 (33)	0.35 (28)	0.28 (11)	0.22 (48)	0.38 (18)	0.22 (37)	0.29 (22)	0.15 (31)
YRI	0.18 (6)	0.21 (186)	0.25 (33)	0.25 (32)	0.21 (11)	0.22 (44)	0.37 (18)	0.21 (38)	0.26 (25)	0.22 (25)

Table S19: Mean values for AT→GC and GC→AT non-synonymous mutations, for those amino-acid changes that can result from both classes of mutations. The values in parantheses represent the sample sizes. Dark green: the difference between mean values was statistically significant (randomization, p-value <0.05), and the mean was higher for AT→GC mutations. Light green: the difference between mean values was not statistically significant (randomization test, p-value >=0.05), and the mean was higher for AT→GC mutations. Light red: the difference between mean values was not statistically significant (randomization test, p-value >=0.05), and the mean was lower for AT→GC mutations. No cases were found where the mean is significantly lower for AT→GC mutations.

References

Duret, L., and Arndt, P. F. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*, **4**(5), e1000071.

Frazer, Kelly A, Ballinger, Dennis G, Cox, David R, Hinds, David A, Stuve, Laura L, Gibbs, Richard A, Belmont, John W, Boudreau, Andrew, Hardenbol, Paul, Leal, Suzanne M, Pasternak, Shiran, Wheeler, David A, Willis, Thomas D, Yu, Fuli, Yang, Huanming, Zeng, Changqing, Gao, Yang, Hu, Haoran, Hu, Weitao, Li, Chaohua, Lin, Wei, Liu, Siqi, Pan, Hao, Tang, Xiaoli, Wang, Jian, Wang, Wei, Yu, Jun, Zhang, Bo, Zhang, Qingrun, Zhao, Hongbin, Zhao, Hui, Zhou, Jun, Gabriel, Stacey B, Barry, Rachel, Blumenstiel, Brendan, Camargo, Amy, Defelice, Matthew, Faggart, Maura, Goyette, Mary, Gupta, Supriya, Moore, Jamie, Nguyen, Huy, Onofrio, Robert C, Parkin, Melissa, Roy, Jessica, Stahl, Erich, Winchester, Ellen, Ziaugra, Liuda, Altshuler, David, Shen, Yan, Yao, Zhijian, Huang, Wei, Chu, Xun, He, Yungang, Jin, Li, Liu, Yangfan, Shen, Yayun, Sun, Weiwei, Wang, Haifeng, Wang, Yi, Wang, Ying, Xiong, Xiaoyan, Xu, Liang, Waye, Mary M Y, Tsui, Stephen K W, Xue, Hong, Wong, J. Tze-Fei, Galver, Luana M, Fan, Jian-Bing, Gunderson, Kevin, Murray, Sarah S, Oliphant, Arnold R, Chee, Mark S, Montpetit, Alexandre, Chagnon, Fanny, Ferretti, Vincent, Leboeuf, Martin, Olivier, Jean-Franois, Phillips, Michael S, Roumy, Stphanie, Salle, Clmentine, Verner, Andrei, Hudson, Thomas J, Kwok, Pui-Yan, Cai, Dongmei, Koboldt, Daniel C, Miller, Raymond D, Pawlikowska, Ludmila, Taillon-Miller, Patricia, Xiao, Ming, Tsui, Lap-Chee, Mak, William, Song, You Qiang, Tam, Paul K H, Nakamura, Yusuke, Kawaguchi, Takahisa, Kitamoto, Takuya, Morizono, Takashi, Nagashima, Atsushi, Ohnishi, Yozo, Sekine, Akihiro, Tanaka, Toshihiro, Tsunoda, Tatsuhiko, Deloukas, Panos, Bird, Christine P, Delgado, Marcos, Dermitzakis, Emmanouil T, Gwilliam, Rhian, Hunt, Sarah, Morrison, Jonathan, Powell, Don, Stranger, Barbara E, Whittaker, Pamela, Bentley, David R, Daly, Mark J, de Bakker, Paul I W, Barrett, Jeff, Chretien, Yves R, Maller, Julian, McCarroll, Steve, Patterson, Nick, Pe'er, Itsik, Price, Alkes, Purcell, Shaun, Richter, Daniel J, Sabeti, Pardis, Saxena, Richa, Schaffner, Stephen F, Sham, Pak C, Varilly, Patrick, Altshuler, David, Stein, Lincoln D, Krishnan, Lalitha, Smith, Albert Vernon, Tello-Ruiz, Marcela K, Thorisson, Gudmundur A, Chakravarti, Aravinda, Chen, Peter E, Cutler, David J, Kashuk, Carl S, Lin, Shin, Abecasis, Gonalo R, Guan, Weihua, Li, Yun, Munro, Heather M, Qin, Zhaohui Steve, Thomas, Daryl J, McVean, Gilean, Auton, Adam, Bottolo, Leonardo, Cardin,

Niall, Eyheramendy, Susana, Freeman, Colin, Marchini, Jonathan, Myers, Simon, Spencer, Chris, Stephens, Matthew, Donnelly, Peter, Cardon, Lon R, Clarke, Geraldine, Evans, David M, Morris, Andrew P, Weir, Bruce S, Tsunoda, Tatsuhiko, Mullikin, James C, Sherry, Stephen T, Feolo, Michael, Skol, Andrew, Zhang, Houcan, Zeng, Changqing, Zhao, Hui, Matsuda, Ichiro, Fukushima, Yoshimitsu, Macer, Darryl R, Suda, Eiko, Rotimi, Charles N, Adebamowo, Clement A, Ajayi, Ike, Aniagwu, Toyin, Marshall, Patricia A, Nkwodimmah, Chibuzor, Royal, Charmaine D M, Leppert, Mark F, Dixon, Missy, Peiffer, Andy, Qiu, Renzong, Kent, Alastair, Kato, Kazuto, Niikawa, Norio, Adewole, Isaac F, Knoppers, Bartha M, Foster, Morris W, Clayton, Ellen Wright, Watkin, Jessica, Gibbs, Richard A, Belmont, John W, Muzny, Donna, Nazareth, Lynne, Sodergren, Erica, Weinstock, George M, Wheeler, David A, Yakub, Imtaz, Gabriel, Stacey B, Onofrio, Robert C, Richter, Daniel J, Ziaugra, Liuda, Birren, Bruce W, Daly, Mark J, Altshuler, David, Wilson, Richard K, Fulton, Lucinda L, Rogers, Jane, Burton, John, Carter, Nigel P, Clee, Christopher M, Griffiths, Mark, Jones, Matthew C, McLay, Kirsten, Plumb, Robert W, Ross, Mark T, Sims, Sarah K, Willey, David L, Chen, Zhu, Han, Hua, Kang, Le, Godbout, Martin, Wallenburg, John C, L'Archevque, Paul, Bellemare, Guy, Saeki, Koji, Wang, Hongguang, An, Daochang, Fu, Hongbo, Li, Qing, Wang, Zhen, Wang, Renwu, Holden, Arthur L, Brooks, Lisa D, McEwen, Jean E, Guyer, Mark S, Wang, Vivian Ota, Peterson, Jane L, Shi, Michael, Spiegel, Jack, Sung, Lawrence M, Zacharia, Lynn F, Collins, Francis S, Kennedy, Karen, Jamieson, Ruth, and Stewart, John. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**(7164), 851–861.

Lohmueller, K. E., Indap, A. R., Schmidt, S., Boyko, A. R, Hernandez, R D, Hubisz, M J, Sninsky, J J, White, T J, Sunyaev, S R, Nielsen, R, Clark, A G, and Bustamante, C D. 2008. Proportionally more deleterious genetic variation in European than in African populations. *Nature*, **451**(7181), 994–997.

HapMap sample	Intergenic		Intron		Synonymous		NonSynonymous	
	Low GC	High GC	Low GC	High GC	Low GC	High GC	Low GC	High GC
CEU	0.007	0.017	0.016	0.022	0.036	0.045	0.014	0.01
CHB+JPT	0.008	0.016	0.015	0.021	0.041	0.076	-0.007	0.059
YRI	0.009	0.017	0.021	0.025	-0.001	0.063	0.024	0.024

Table 1: Δd values for different GC-contents. Each set of SNPs was divided into two equal-size classes, according to the GC-content in 100bp flanking regions. The regions of low and high recombination are defined based on the distance from SNPs to recombination hotspots.

HapMap sample	Intergenic		Intron		Synonymous		NonSynonymous	
	Low CpG	High CpG	Low CpG	High CpG	Low CpG	High CpG	Low CpG	High CpG
CEU	0.009	0.018	0.019	0.023	0.047	0.038	0.027	0.01
CHB+JPT	0.01	0.017	0.019	0.02	0.064	0.057	0.01	0.022
YRI	0.012	0.019	0.024	0.028	0.026	0.038	0.016	0.05

Table 2: Δd values for different CpG contents. Each set of SNPs was divided into two equal-size classes, according to the CpG observed/expected ratio in 100bp flanking regions. The regions of low and high recombination are defined based on the distance from SNPs to recombination hotspots.